Lazar Tenjović

STATISTIKA U PSIHOLOGIJI

priručnik

Lazar Tenjović

STATISTIKA U PSIHOLOGIJI

PRIRUČNIK

Dopunjeno izdanje

DODATAK 2

–

DODATAK 4

(1-10)
6 - 17 str.

Beograd

PREDGOVOR

Priručnik je prvenstveno namenjen studentima psihologije na Filozofskom fakultetu u Beogradu, kao osnovno štivo za vežbe iz predmeta *Statistika u psihologiji*, a može biti koristan i diplomiranim psiholozima i istraživačima kao podsetnik pri statističkoj analizi podataka. U Priručniku su prikazani osnovni statistički postupci koji se najčešće koriste za analizu podataka psiholoških istraživanja. Osim sažetog prikaza teorijske i matematičke osnove svakog postupka, dato je i objašnjenje načina na koji se statistička metoda može primeniti pomoću programa SPSS za Windows. Delovi teksta u kojima je objašnjen postupak primene date metode u programu SPSS sadrže i primere ispisa iz ovog programa, kao i komentare ispisa. Ovi delovi su od ostalog dela teksta odvojeni redovima zvezdica.

Ovo dopunjeno izdanje Priručnika sadrži osnovni tekst koji je identičan onome u prvom izdanju. Pored osnovnog teksta novo izdanje Priručnika sadrži četiri dodatka: prva dva su postojala i u prvom izdanju, dok su Dodatak 3 i 4 novi. U Dodatku 1 objašnjene su osnove rada na PC računaru u Windows okruženju, dok Dodatak 2 sadrži sažet prikaz osnovnih znanja iz teorije verovatnoće i matematičke statistike. Dodatak 3 sadrži veoma kratak i sažet prikaz osnovnih pojmova iz matrične algebre. U Dodatku 4 data je Tabela sa ispravkama krupnijih grešaka koje su prisutne u osnovnom tekstu Priručnika. Osim toga, Dodatak 4 sadrži veoma važne napomene i dopune u vezi sa pojedinim delovima osnovnog teksta Priručnika. Budući da bi intervencije u samom osnovnom tekstu znatno podigle inače nepopularnu cenu Priručnika, odlučili smo da u ovom izdanju ostavimo osnovni tekst netaknut, a da ispravke i dopune osnovnog teksta dodamo na kraju, što je i učinjeno u Dodatku 4. Stoga preporučujem čitaocu da pre korišćenja osnovnog teksta sam unese ispravke koje se nalaze u tabeli na početku Dodatka 4 i da određenim znakom u osnovnom tekstu naznači na kojim mestima treba konsultovati Dodatak 4. Korisnici prethodnog izdanja Priručnika mogu Dodatke 3 i 4 skinuti sa internet prezentacije predmeta Statistika u psihologiji, na URL adresi: http://www.f.bg.ac.yu/psihologija/statistika/prirucnik.htm

Osim tekstova koji su navedeni kao literatura na kraju knjige, najvažniji izvor informacija koje su korišćene u pisanju Priručnika bile su beleške sa predavanja prof. **Konstantina Momirovića** koja su držana 90-tih godina studentima psihologije na Filozofskom fakultetu u Beogradu. Predavanja prof. Momirovića podrazumevala su prethodno savladan kurs osnovnih pojmova iz Matematike (Linearne algebre i Osnova teorije verovatnoće posebno) koja je istim studentima držao prof. **Mile Bogdanović**. Budući da su predavanja prof. Momirovića i Bogdanovića, koja sam imao sreću da slušam, presudno uticala na sadržaj Priručnika, neophodni uslov za njegovo praćenje je poznavanje osnova Teorije verovatnoće i Linearne algebre. Korisnicima Priručnika koji ne raspolažu osnovnim znanjima iz ovih oblasti osim Dodatka 2 i Dodatka 3 na kraju Priručnika, veoma koristan može biti i dodatak Elementi linearne algebre u knjizi K. Momirovića, B. Wolfa i D. A. Popovića *Uvod u teoriju merenja I*. Zbog obimnosti materijala pojedini delovi teksta su napisani veoma sažeto. Kako bi se ovi delovi Priručnika mogli nesmetano pratiti, treba pri njegovom korišćenju uporedo koristiti knjige *Statistika za psihologe* Č. Dragićevića i *Multivarijaciona analiza* Zlatka Kovačića. Potpuni bibliografski podaci o knjigama koje su navedene u prethodnom delu teksta dati su u spisku literature.

Veliku zahvalnost za pomoć u pisanju pojedinih delova teksta i doterivanju celog Priručnika dugujem kolegi **Aleksandru Zoriću**.

Sugestije i primedbe korisnici Priručnika mogu slati na elektronsku adresu: ltenjovi@f.bg.ac.yu

Лазар Тењовић

S a d r ž a j :

I UVOD U PROGRAMSKI PAKET SPSS FOR WINDOWS 1

I.1. Komponente programskog paketa SPSS for Windows 1

I.2. Aktiviranje programa SPSS i izlazak iz programa 1

II ORGANIZACIJA I UNOS PODATAKA 5

II.1. Unos podataka u SPSS-ovom editoru podataka 5

II.2. Učitavanje ASCII podataka, prethodno unetih u nekom od. standardnih editora, u prozor za podatke programa SPSS 8

II.3. Učitavanje podataka unetih u standardnim bazama podataka u prozor za podatke programa SPSS 9

II.4. Učitavanje podataka koji su već sačuvani u formatu SPSS-a za Windows u prozor za podatke 9

III OSNOVNI POJMOVI I OZNAKE ČIJE POZNAVANJE JE NEOPHODNO ZA RAZUMEVANJE DALJEG TEKSTA 10

IV SREĐIVANJE PODATAKA 18

IV.1. Uređivanje podataka 19

IV.2. Klasifikovanje podataka - pravljenje raspodele učestalosti 19

IV.3. Transformisanje unetih vrednosti varijable 21

IV.4. Mere centralne tendencije 25

IV.5. Mere varijabilnosti 28

IV.6. Percentili i percentilni rangovi 32

IV.7. Momenti i momentni koeficijenti jedne kvantitativne varijable 33

IV.8. Neke važne transformacije podataka na kvantitativnoj varijabli 35

IV.9. Grafičko prikazivanje jedne kvantitativne varijable 37

V STATISTIČKI OPIS I GRAFIČKI PRIKAZ JEDNE KATEGORIČKE (NOMINALNE) VARIJABLE 43

V.1. Definisanje i kodiranje nominalne (kategoričke) varijable 43

V.2. Statistički opis kategoričke varijable 44

V.3. Grafičko predstavljanje kategoričke varijable 46

VI STATISTIČKO ZAKLJUČIVANJE 49
VI.1. Statističko ocenjivanje parametara 49
VI.2. Testiranje statističkih hipoteza 56

VII TESTIRANJE HIPOTEZE O NORMALNOSTI RASPODELE VARIJABLE U POPULACIJI 59

VIII TESTIRANJE HIPOTEZE O ARITMETIČKOJ SREDINI POPULACIJE 63

IX TESTIRANJE ZNAČAJNOSTI RAZLIKA IZMEĐU ARITMETIČKIH SREDINA - t test za nezavisne i zavisne uzorke 65

X ANALIZA VARIJANSE 71
X.1. Jednofaktorijalna univarijatna analiza varijanse 71
X.2. Dvofaktorijalna univarijatna analiza varijanse 88

XI LINEARNA KORELACIJA I JEDNOSTRUKA REGRESIJA 103
XI.1. Linearna korelacija 103
XI.2. Jednostruka linearna regresija 112

XII HI-KVADRAT TEST: TESTIRANJE HIPOTEZA O DISTRIBUCIJI VEROVATNOĆA KATEGORIČKE VARIJABLE I ANALIZA KONTINGENCIJSKIH TABELA 123
XII.1. Testiranje hipoteze o distribuciji verovatnoća kategoričke varijable ... 123
XII.2. Analiza kontingencijskih tabela - testiranje nezavisnosti dveju kategoričkih varijabli 125
XII.3. Koeficijenti asocijacije dveju kategoričkih varijabli 128

XIII NEPARAMETRIJSKE STATISTIČKE METODE (METODE BEZ STROGIH PRETPOSTAVKI O OBLIKU DISTRIBUCIJE MERA U POPULACIJI) ZA RANGOVANE PODATKE 134
XIII.1. Parametrijske i "neparametrijske" metode 134
XIII.2. Neparametrijske metode korelacije rangova 135
XIII.3. Neparametrijske metode za testiranje pretpostavke da uzorci pripadaju populacijama sa istim distribucijama ili populacijama sa istim centralnim vrednostima 141

XIV MULTIPLA REGRESIONA ANALIZA 150
XIV.1. Pojam kompozitne (složajne) varijable 150
XIV.2. Multipla regresiona analiza 151
XIV.3. Koeficijenti parcijalne i semiparcijalne korelacije 165

XV ANALIZA GRUPISANJA 169

XVI KANONIČKA KORELACIONA ANALIZA 188

XVII KANONIČKA DISKRIMINACIONA ANALIZA 200

XVIII ANALIZA GLAVNIH KOMPONENTI 219

XIX ANALIZA TRENDA 234
XI.1. Pojam vremenskih serija 234
XI.2. Analiza polinomskih trendova 234

DODATAK 1.: OSNOVE RADA NA PC RAČUNARU POD OPERATIVNIM SISTEMOM WINDOWS I OSNOVE KORIŠĆENJA INTERNET EKSPLORERA

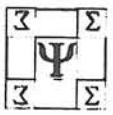
DODATAK 2.: OSNOVE TEORIJE VEROVATNOĆE I MATEMATIČKE STATISTIKE NEOPHODNE ZA RAZUMEVANJE STATISTIČKOG ZAKLJUČIVANJA

DODATAK 3.: OSNOVNI POJMOVI MATRIČNE ALGEBRE

DODATAK 4.: ISPRAVKE KRUPNIJIH GREŠAKA U PRIRUČNIKU I VAŽNE NAPOMENE I DOPUNE U VEZI SA POJEDINIM DELOVIMA OSNOVNOG TEKSTA

Literatura

STATISTIKA U PSIHOLOGIJI



PRIRUČNIK

"Statistiku je mnogo lakšé naučiti nego (p)ostati psiholog bez nje."

"... zadatak korisnika statistike je da izabere metod za koji nezadovoljenost uslova ima najmanje ozbiljne posledice... Vešt primenjeni statističar je onaj koji pravi najbolji izbor između loših alternativa."

D.W. Stilson

I. UVOD U PROGRAMSKI PAKET SPSS for WINDOWS

I.1. Komponente programskog paketa SPSS for Windows

Programski paket **SPSS** (Statistical Package for Social Sciences) **for WINDOWS** ima nekoliko komponenti:

1. **Programe** koji: učitavaju podatke, izvode zadate analize i daju odgovarajući ispis rezultata. Programi se (po pravilu) nalaze u direktorijumu (folderu) C:\SPSSWIN.

2. **Fajlove sa podacima** koji su sačuvani u SPSS za WINDOWS formatu i imaju podrazumevanu (default) ekstenziju **SAV**.

3. **Fajlove sa ispisima** rezultata koji imaju podrazumevanu ekstenziju **LST** (u starijim verzijam programa) ili **SPO** (u novijim verzijama).(Fajlovi sa graficima napravljenim u SPSS programu imaju podrazumevanu ekstenziju CHT).

4. **Komandne**, tj. **sintaksne fajlove** sa podrazumevanom ekstenzijom **SPS.**[1]

I.2. Aktiviranje programa SPSS i izlazak iz programa

Windows 3.xx:[2] Na odzivni znak DOS-a, komandom **WIN** <ENTER>, ulazi se u WINDOWS okruženje a potom dva puta klikne na sličicu, tj. ikonicu SPSS u programskoj grupi sličnog naziva. Iz programa se izlazi bilo dvostrukim klikom na kontrolno sanduče (control menu box) u gornjem levom uglu glavnog prozora **SPSS** programa, bilo odabirom SPSS-ovog menija **File**, a potom opcije **Exit**.

Windows95:[3] Po pritisku na dugme **Start** (na liniji poslova), odabere se stavka **Programs** i u okviru programske grupe **SPSS** klikom na stavku **SPSS** aktivira program. Iz programa se izlazi bilo klikom na znak dugme **x** za zatvaranje prozora (u gornjem desnom uglu), ili odabirom komandi **File** pa **Exit** iz menija programa.

Kada se program aktivira obično se otvara nekoliko prozora:

1. **Programski ili aplikacioni prozor** (Slika 1), čiji je naslov **SPSS for Windows**, a u okviru kojega je **MENI** programa (sa rečima od **File** do **Help**). Na donjem rubu ovog programskog prozora ispisuju se poruke o trenutnom statusu programa i druge važne poruke (npr. o tome da li je uključen filter za odabir samo nekih ispitanika za analizu i sl.).

---

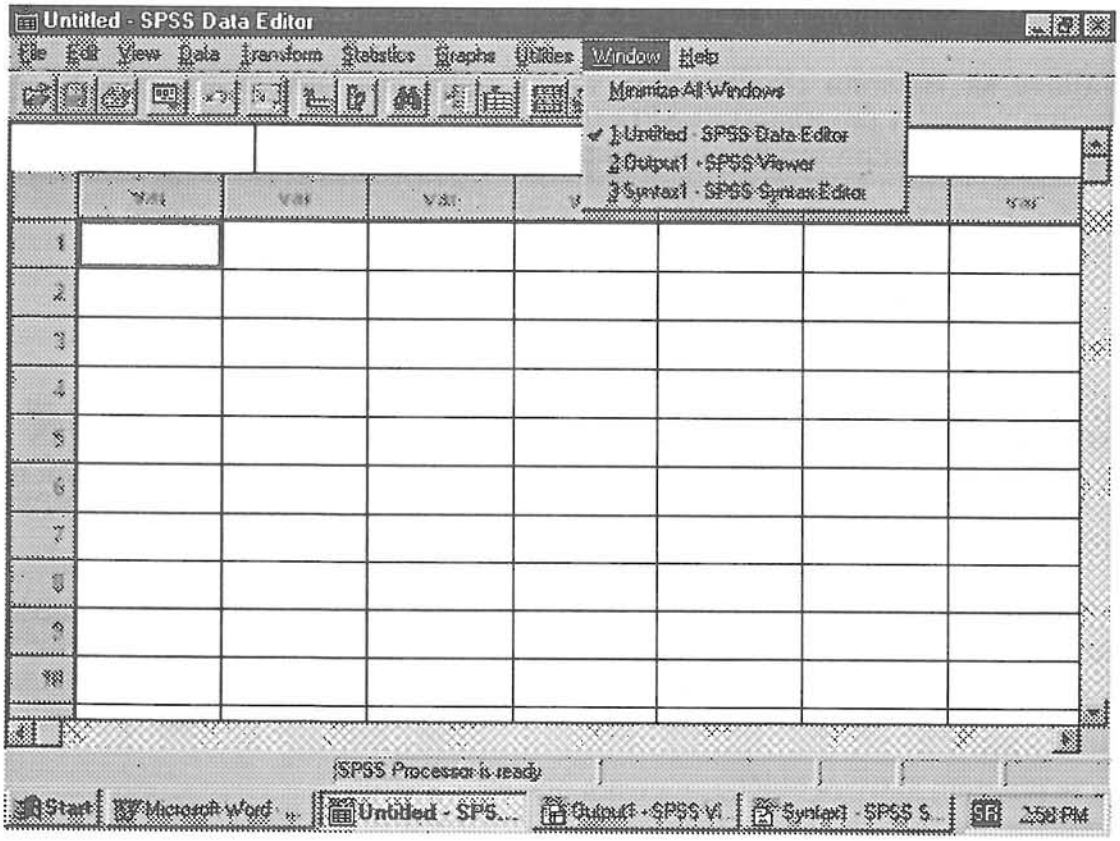
[1] Sintaksa komandi, koje se (iz menija ili iz sintaksnog prozora) zadaju programu da ih izvrši, čuva se u fajlu SPSS.JNL (žurnal fajl) koji se obično nalazi u direktoriju C:\WINDOWS\TEMP. Ako je u meniju **Edit Preferences** (u starijim verzijama) ili **View Options** (u novijim verzijama) programa SPSS, u okviru **Session journal**, aktivirano snimanje sintakse u fajl (**Record syntax in journal**) i odabrana opcija **Append** onda će SPSS.JNL fajl brzo narasti do nepodnošljive veličine. Mudro bi bilo, u tom slučaju, u nekom od editora povremeno prazniti sadržaj ovog fajla da se ne bi desilo da jednog dana zauzme ceo prostor na harddisku!

[2] U ovom tekstu umesto određenog broja pojedinih verzija Windows-a 3 stajaće xx.

[3] Praktično sva objašnjenja za rad pod W95 važe i za rad pod Windows98 ili WindowsNT sistemima.

2. **Prozor za ispis** rezultata analiza i poruka o greškama (slika 2). Naslov ovog prozora je **!OUTPUTn** (n je određeni broj) ili **!OUTPUTn SPSS Viewer**, ili putokaz i ime fajla u kojem je pohranjen sadržaj ispisa. Znak **!** služi za određivanje (designate) jednog prozora za ispis kao aktivnog kada je više prozora za ispis otvoreno. Naime, novi ispisi rezultata će automatski ići u unapred određeni prozor za ispis, tj. onaj prozor čiji je znak usklika prikazan masnim fontom. Klikom na uskličnik unutar određenog prozora za ispis dati prozor se predodređuje za naredni ispis).

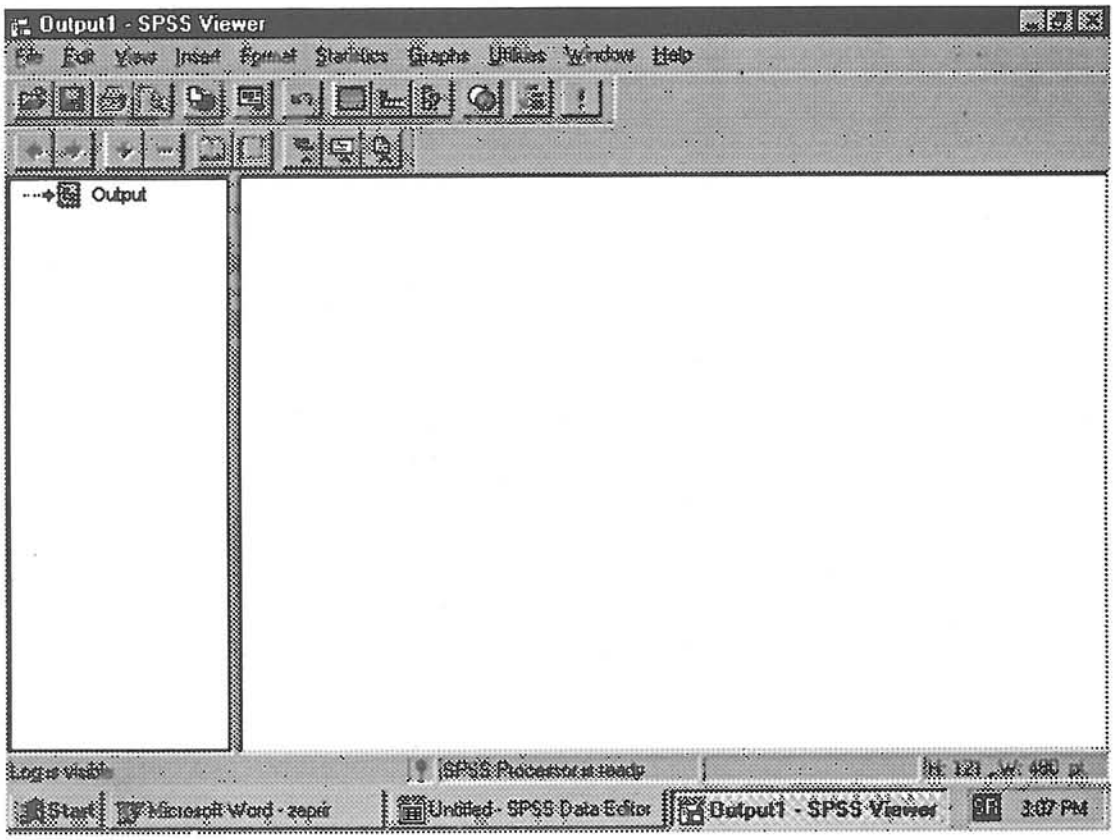
Slika 1: Glavni prozor (sa tabelom za podatke) u programu SPSS, verzija 8.0



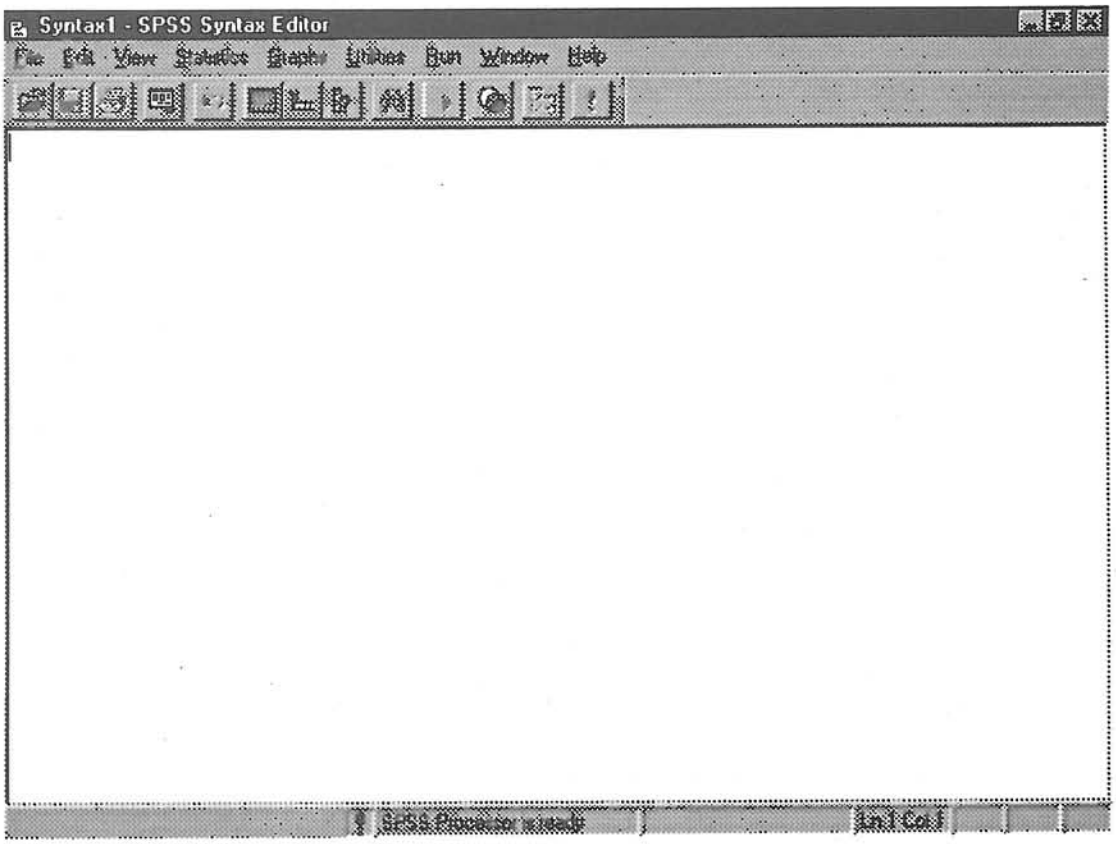
3. **Prozor za podatke**. Ukoliko je reč o novom fajlu u koji tek treba uneti podatke naslov prozora sa tabelom za podatke će biti **NEWDATA** ili **UNTITLED SPSS Data Editor**. Ukoliko se učitavaju već uneti podaci naslov prozora za podatke će imati **PUTOKAZ** do direktorijuma u kojem je fajl i **IME FAJLA** sa podacima.

4. **Sintaksni prozor** (slika 3) čiji je naslov **!SYNTAXn** (n je određeni broj) ili **!SYNTAXn SPSS Syntax Editor**, tj. prozor u koji se mogu interaktivno unositi (ili učitavati unapred pripremljene) komande za analizu podataka. /Ovaj način rada namenjen je pre svega onima koji su dobro ovladali sintaksom komandnog jezika SPSS programa/. U starijim verzijama Windows-a unutar ovog prozora su stavke menija za rad sa komandama (**Run**- za izvršenje odabrane komande, **Syntax**- za podsećanje na sintaksu komandi, **!** - za definisanje aktivnog, tj. predodređenog /designate/ sintaksnog prozora). Pojava sintaksnog prozora pri startovanju programa nije obavezna već je opciona i pojavljuje se samo ukoliko je u definisanju rada programa eksplicitno odabrana (u meniju **Edit Preferences** ili **View Options**).

Slika 2: Prozor za ispis



Slika 3: Sintaksni prozor



Svi prozori SPSS-a imaju iste bitne elemente kao i ostali prozori WINDOWS okruženja i ista osnovna pravila o smanjenju, povećavanju, premeštanju, uređivanju, pregledanju sadržaja, zatvaranju prozora, prelaska iz prozora u prozor i sl.[4] O radu sa prozorima u Windows okruženju može se pogledati u **Dodatku 1** ovog priručnika.

[4] Prozor sa podacima (u starijim verzijama programa i jedan prozor za ispis) ne može se zatvoriti dokle god program radi. On se zatvara automatski pri napuštanju programa.

II. ORGANIZACIJA I UNOS PODATAKA

Podaci za analizu pomoću SPSS programa organizuju se u tabelu, tj. matricu reda **n** x **m** u kojoj su redovi entiteti, slučajevi, ispitanici ili jedinice posmatranja, a kolone atributi, svojstva ovih entiteta, tj. varijable. Podaci se mogu unositi direktno u SPSS tabelu, tj. editor podataka. Isto tako, podaci koji su uneti u bilo kojem standardnom editoru ili procesoru teksta (i sačuvani u ASCII - American Standard Code for Information Interchange- formatu), mogu se učitati u ovu tabelu (uz neophodne dodatne definicije). Takođe, podaci iz standardnih WINDOWS baza podataka (Exel, Dbase, Access...) mogu se bez dodatnih definicija učitati i analizirati SPSS programom.

II.1.Unos podataka u SPSS-ovom editoru podataka

Tabela Editora podataka u koju se prvi put unose podaci podeljena je u redove koji su označeni brojevima od 1 pa nadalje i kolone čiji naslovi su var0001 do var0000m. Mada to nije nužan sled operacija najbolje je prethodno definisati varijable pa potom unositi podatke u ovu tabelu.

Definisanje varijabli:

- Kliknuti mišem na naslov prve kolone VAR (slika 4).

- U glavnom meniju programa odabrati stavku **Data** a potom iz "padajućeg" menija opciju **Define Variable.**
(Prethodne dve operacije mogu se zameniti levim duplim klikom miša na naslov kolone var0000m)

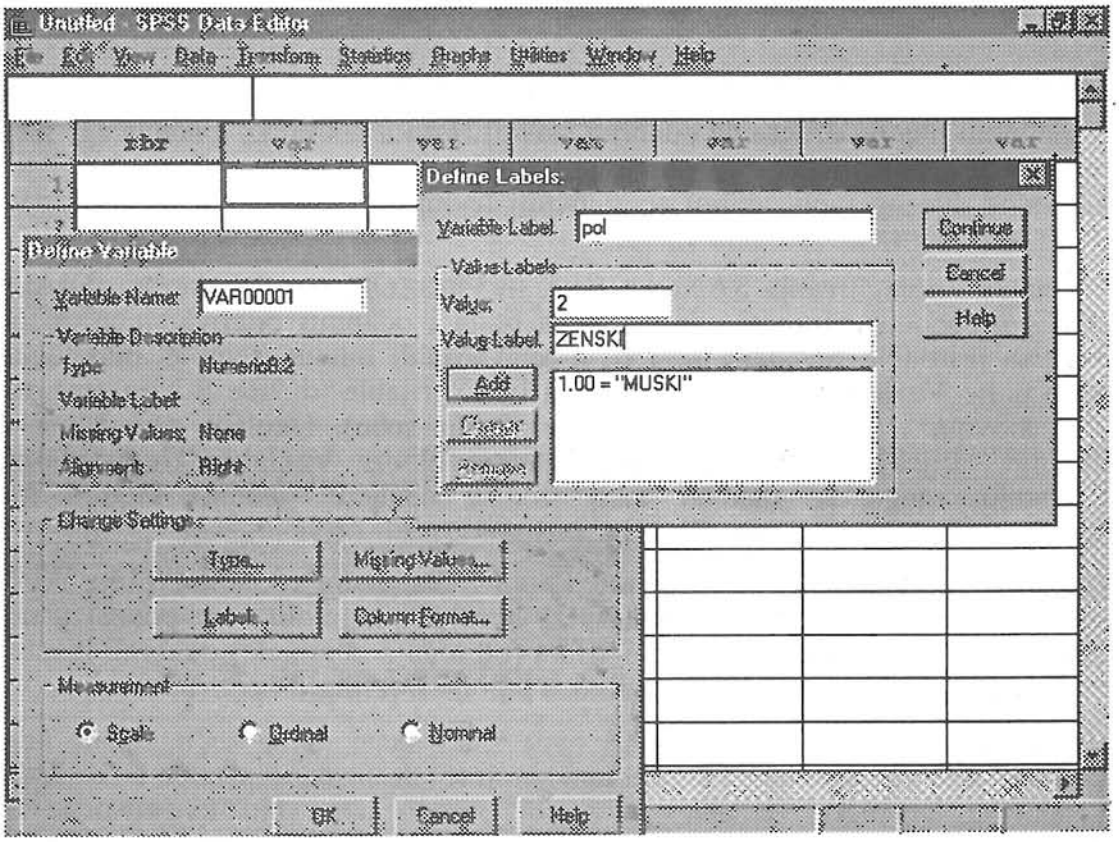
- U trenutku pojavljivanja SANDUČETA ili PROZORA za DIJALOG (DIALOG BOX) potrebno je popuniti odgovarajuća polja i odabrati odgovarajuće opcije u skladu sa prirodom podataka koji će u toj varijabli biti uneseni. U pravougaono ulazno polje **Variable name** treba upisati **ime** varijable (dužine 8 karaktera, mora početi slovom i ne sme imati praznine unutar sebe). Ukoliko podrazumevane definicije varijable koje se pojavljuju unutar okvira **Variable Description** ne odgovaraju, mogu se menjati tako što se odabere (jednim klikom miša) odgovarajuće dugme u okviru **Change Settings**:

Type- služi za definisanje tipa varijable (Numerička, String...), širine (polje **Width**), broja decimalnih mesta (polje **Decimal Places**). Npr. ukoliko su podaci u nekoj varijabli numerički i imaju najviše 6 cifara (od toga 3 decimale) odabrali bismo tip **Numeric**, u polje **Width** bismo upisali 6, a u polje **Decimal Places** 3.

Zapamtite: Numerički tip varijable ovde znači samo to da se u kolonu koja predstavlja tu varijablu mogu unositi samo brojevi i decimalni separator, a ne i drugi karakteri. Numeričke varijable po svojoj prirodi mogu biti kvantitativne ili kategoričke (kvalitativne). O značenju cifara u numeričkim varijablama (da li su cifre oznake kvantiteta ili kvaliteta) mora voditi računa sam korisnik programa! Varijable tipa String ne mogu biti korišćene u računanju! Stoga je preporučljivo i kvalitativne (tj. kategoričke) varijable kodirati numerički. Na primer, bolje je (i lakše) u varijablu POL unositi cifre 1 i 2 (kao oznake za muškarce i žene) nego unositi reči MUŠKI i ŽENSKI . Ukoliko su kategorije neke kategoričke varijable ipak unete tekstualno one se mogu prevesti u numerički oblik odabirom menija **Transform** i potom opcije **Automatic Recode**.

Labels - služi za određivanje naziva varijable i naziva pojedinih vrednosti varijable. **Naziv** varijable (**Variable labels**) nema nužno nikakve veze sa **imenom** varijable (**Variable name**). Program za sve analize koristi ime varijable (i ono je ograničene dužine- do 8 karaktera), a nazivi varijabli i nazivi pojedinih vrednosti varijabli mogu biti znatno duži i koriste pre svega za snalaženje u ispisima rezultata. U dijalogu **Labels** može se popuniti polje **Variable Labels** željenim nazivom varijable. Npr. varijabla čije je ime KMS može dobiti naziv SKOR NA KRAUN MARLOVOJ SKALI. U polju **Value** dijaloga **Value Labels** unose se pojedine vrednosti varijable koje se pojavljuju u podacima, a u polje **Value Labels** unosi se željeni naziv ove vrednosti. Npr. u varijabli POL vrednost 1 koja označava muškarce može dobiti naziv MUŠKI, a vrednost 2 ŽENSKI. Po definisanju vrednosti varijable i njenog naziva potrebno je kliknuti na reč **Add** i ova definicija naziva vrednosti će biti uneta u pravougaoni prostor na dnu dijalog prozora. Ukoliko se neka od već unetih definicija menja ona se označava tj. markira, ispravi šta je potrebno u poljima **Value** i **Value Label**, a potom klikne na reč **Change**. Izbacivanje definicije naziva vrednosti vrši se tako da se odabere data definicija i potom klikne na dugme **Remove**.

Slika 4: Prozor za dijalog za definisanje varijabli sa poddijalogom za naziv varijable i nazive vrednosti na varijabli



Missing value - služi za definisanje nedostajućih podataka. Ukoliko za nekog ispitanika nemamo podatak na nekoj varijabli moguće je u njegovoj ćeliji za tu varijablu ostaviti prazninu. Ovu će prazninu SPSS shvatiti kao **sistemsku nedostajuću vrednost (SYSMIS)**. Međutim, ukoliko iz bilo kojih razloga unosimo određeni broj ili karakter kao znak da za ispitanika na datoj varijabli nemamo podatak, to moramo definisati u dijalogu **Missing value** za datu varijablu.

Ovako definisane nedostajuće vrednosti zovu se **User Missing value**, tj. korisničke nedostajuće vrednosti.

Column Format- služi za definisanje formata ispisa varijable u tabeli. Ova definicija nema nikakvog uticaja na format zapisivanja podataka već samo na način na koji će podaci biti prikazani.

> **Zapamtite**: Unutar svakog PROZORA ZA DIJALOG po završetku posla treba odabrati opciju **Continue** ili **OK**, ukoliko je definicija u redu a **Cancel** ukoliko želimo da iz prozora za dijalog izađemo, a da se definicija koju smo počeli ne uzme u obzir.

Opcije u okviru **Measurement** (slika 4) u novijim verzijama programa omogućuju da se definiše merna skala (o mernim skalama se može pročitati u glavi **III** ovog Priručnika) sa koje dolaze podaci na datoj varijabli. Uključena opcija **Scale** u tom okviru znači da su podaci na varijabli bar intervalnog tipa, opcija **Ordinal** govori da je reč o rangovima, a ukoliko je uključena opcija **Nominal** znači da je reč o kategoričkoj varijabli. Istina, praktično jedina korist od definisanja prave opcije u okviru **Measurement** je u tome što će program odbiti da napravi grafičke prikaze koji su predviđeni za intervalne varijable na varijablama kategoričkog tipa. Bar zasad, program uopšte ne reaguje ako se primeni neodgovarajuća statistička procedura za dati tip varijable!

- Kada je definisanje varijable gotovo, klikom na dugme **OK** unutar PROZORA za DIJALOG proces se završava i unutar tabele editora pojaviće se u odgovarajućoj koloni novo ime varijable. Postupak se potom ponavlja za preostale varijable. Jasno je da se na ovaj način mogu menjati definicije ranije definisanih varijabli. Tako se u prvoj koloni matrice podataka na slici 4 nalazi varijabla RBR koja je već definisana, a u drugoj koloni će biti varijabla pol kada njeno definisanje bude gotovo na osnovu dijaloga koji je prikazan na slici 4.

Unos podataka

Kada je definisanje svih varijabli završeno, klikom na prvu ćeliju unutar tabele omogućen je unos podataka. Po unošenju podatka u datu ćeliju, strelicama za pomeranje kursora prelazi se ili u sledeći red ili u sledeću kolonu zavisno od toga da li se podaci unose po ispitanicima (što je najčešći slučaj) ili po varijablama.
Povremeno je potrebno (a na kraju unosa svih podataka obavezno) da se, u trenutku kada je Prozor za podatke aktivni prozor, podaci trajno zapišu naredbom **File Save Data**. Pri prvom zapisivanju fajla sa podacima potrebno je odabrati komandu **File Save As** i dati ime fajlu (poželjno je da ostane podrazumevana ekstenzija SAV).[5] Ukoliko je neophodno da se novija verzija starog fajla sa podacima sačuva pod drugim imenom, treba u meniju **File** odabrati opciju **Save as** i dati novo ime fajlu sa podacima.[6]

---

[5] Podaci iz SPSS editora mogu se sačuvati i u drugim formatima. U ovakvim slučajevima potrebno je u meniju **File** odabrati opciju **Save as** a potom odabrati format izlaznih podataka (u prozoru **Save file as type**) i dati odgovarajuće ime fajlu.

[6] Na desnoj i donjoj ivici prozora za podatke nalaze se klizači (scroll bar) kojima se može pregledati onaj sadržaj tabele koji nije stao u trenutnu veličinu prozora.

II.2. Učitavanje ASCII podataka, prethodno unetih u nekom od standardnih editora, u prozor za podatke programa SPSS

Objašnjenje ovog postupka daćemo samo za ASCII fajlove sa podacima u tzv. **fiksnom formatu**, tj. za fajlove u kojima se uvek u istoj koloni kod svih ispitanika nalazi podatak za istu varijablu. Postoji i tzv. slobodni format podataka ali se on ređe koristi.

Učitavanje ASCII podataka u fiksnom formatu, koji su uneseni u nekom od standardnih editora, u SPSS Editor moguće je uraditi na dva načina:

1. U meniju **File** odabrati opciju **Read ASCII data**. Zatim u DIJALOG prozor treba uneti ime fajla u kojem su podaci ili (ukoliko se fajl nalazi u spisku ponuđenih fajlova) dva puta kliknuti na ime fajla. Potom kliknemo na dugme **Define** i u dijalogu SPSS programa **Define Fixed Variables** definišemo varijable koje treba učitati iz fajla: Ime(a ne Label tj. naziv!) u polju Name; u kojoj koloni je varijabla u izvornom fajlu (**Start Column** i **End Column**), Tip podataka i korisničke nedostajuće vrednosti. Posle definicije svake varijable klikom na reč **Add** prelazimo na sledeću varijablu, a ime varijable koju smo definisali upisuje se u pravouganik sa desne strane.
Ukoliko su podaci za svakog ispitanika uneseni u više od jednog reda u polju **Record** treba pri definisanju varijabli koje nisu u prvom redu vrednost 1 zameniti brojem reda u kojem je varijabla.

2. Drugi način podrazumeva poznavanje sintaksi SPSS komandi. Potrebno je unapred opisati podatke u nekom od editora, tj. komandom **DATA LIST FILE** definisati u kom su fajlu podaci, imena varijabli i kolone u kojima su pojedine varijable.

Na primer komandom:

DATA LIST FILE='a:\nauka\bekrija.dat'/rbr 1-2 pol 3 pijesam 4 stapije 5-6.

definišemo da se fajl sa podacima u ASCII formatu nalazi na disketi koja je u drajvu A i to u direktoriju NAUKA, a zove se bekrija.dat . Osim toga varijabla sa imenom rbr (redni broj) zauzima kolone 1 i 2 , varijabla pol kolonu 3 itd.
Ukoliko je komanda DATA LIST FILE pravilno definisana u meniju SPSS-a potrebno je odabrati opciju File Open Syntax i učitati sintaksni tj. komandni fajl u SYNTAX prozor. Potom se klikne na poziciju ispred reči DATA i držeći pritisnuto levo dugme miša odabere (zatamni) celokupna komanda. Po odabiranju komande DATA LIST FILE miš se otpusti i zatim klikne na dugme Run unutar sintaksnog prozora. U novijim verzijama programa za to se koristi dugme na kojem je oznaka ▸ u toolbaru sintaksnog prozora (slika 3) a, takođe, i komanda **Run** koja se nalazi u meniju sintaksnog prozora (slika 3) i ima nekoliko opcija: **All** - za izvršavanje svih komandi koje se nalaze trenutno u sintaksnom prozoru; **Selection** - za izvršavanje samo selektovanih komandi; **Current** - za izvršavanje komande u kojoj je kurzor i **To End** - za izvršavanje svih komandi od mesta na kojem je kurzor pa do kraja sintaksnog prozora.
Komanda Data list file može se upisati i interaktivno u sintaksni prozor a potom izvršiti njenim odabiranjem i klikom na dugme Run u W3.xx, odnosno tool Run (ikonica sa strelicom nadesno ispod menija). Klik na tool Run u W95 može se zameniti komandom Ctrl R sa tastature (dakle, dok se drži dugme Ctrl pritisne se slovo R).

II.3. Učitavanje podataka unetih u standardnim bazama podataka u prozor za podatke programa SPSS

Po odabiranju stavki menija **File Open Data** u okviru dijaloga **Open Data File** potrebno je definisati u kom je programu pripremljena baza podataka koju treba učitati, tako što se u polju **File Type** odabere odgovarajuća opcija, i, zatim, u polju **File name**, upiše ime fajla (ili odabere iz spiska ukoliko se ime željenog fajla tamo nalazi).

II.4. Učitavanje podataka koji su već sačuvani u formatu SPSS-a za Windows u prozor za podatke

Podaci zapisani u fajlu koji je u formatu SPSS-a za Windows učitavaju se odabirom menija **File Open Data.** Pri pojavi imena ovih fajlova (uobičajeno sa ekstenzijom **.SAV**) treba samo dva puta kliknuti mišem na ime fajla u kojem su podaci. Ukoliko ovog fajla nema u ponuđenom spisku treba odabrati disk na kojem je fajl i direktorijum u kojem se fajl nalazi.

Zapamtite: U jednom trenutku u prozoru sa podacima može biti samo **jedan** fajl sa podacima. Pri učitavanju drugog fajla sa podacima prethodni se uklanja iz radne memorije!

III. OSNOVNI POJMOVI I OZNAKE ČIJE POZNAVANJE JE NEOPHODNO ZA RAZUMEVANJE DALJEG TEKSTA

Entitet e_i

Entiteti - e_i ; i = 1,...,n su objekti, jedinice posmatranja.U psihologiji su to najčešće ljudi, pa se umesto entiteti, u odgovarajućim situacijama, koristi i naziv ispitanici ili subjekti. Oznaka i definiše redni broj, tj. redosled kojim smo entitete uzeli u razmatranje. Ako neki skup entiteta biramo iz šireg skupa entiteta po nekom pravilu onda je reč o **uzorku**.

Skup entiteta[7] E

Skup entiteta: $E= \{e_i; i =1,...,n\} \subset P$. Izraz $E \subset P$ ili, što je identično, $P \supset E$, znači da je skup entiteta deo, tj. **pravi** podskup šireg skupa (populacije)[8].

Populacija P

Postoje vrlo raznorodne definicije populacije. Navešćemo samo neke:

- Populacija je statistička masa, skup varijabilnih elemenata iste vrste. Jasno definisana grupa entiteta ili pojava koje poseduju zajednička obeležja;
- Populacija je skup entiteta na koji želimo da primenimo zaključke istraživanja (Stilson, s.31);
- Populacija je neuređen skup koji sadrži sve entitete koji imaju specifikovana svojstva, tj. takve odlike koje razlikuju članove populacije od entiteta koji to nisu.
- Populacija predstavlja sve entitete za koje vredi neki zakon.

---

[7] Pojam skupa u matematici spada u osnovne pojmove koji se obično ne definišu. Skup se sastoji od elemenata. Ako nema nijednog elementa u skupu takav skup zovemo praznim skupom i označavamo ga simbolom $\varnothing$. Ako su $x_1, x_2,...,x_n$ elementi skupa S, onda se to piše $S = \{ x_1, x_2,...,x_n \}$. Ponekada se ovo piše i kao $S = \{ x | P(x)\}$ čime se definiše da su x-ovi elementi skupa S takvi da imaju svojstvo P. Vertikalna crta posle x se čita kao "pod uslovom da..." ili "takvi da...". Činjenica da je x element skupa S piše se kao $x \in S$ ili $S \ni x$ ("S sadrži x"). Negacija toga, tj. činjenica da x nije element skupa S piše se kao $x \notin S$.

-Dva skupa su identično jednaka ako se sastoje od istih elemenata: $A \equiv B$. Znak $\equiv$ je znak identiteta. Pri tome, redosled elemenata u skupu nije važan. Npr.$\{ 5, 3 ,1 \} \equiv \{ 1, 3, 5 \}$. Skup može biti i jednočlani. Na primer skup može imati samo element S: $\{S\}$ Pri tom $\{S\} \neq S$!

[8] -Ako je svaki element skupa S_1 istovremeno i element skupa S, tj. ako je istinito $(\forall x)\ (x \in S_1 \Rightarrow x \in S)$ onda je skup S_1 deo skupa S. Dakle, $S_1 \subset S$. Znak inkluzije $\subset$ stoji kadgod je S_1 deo skupa S a S sadrži bar jedan element koji nije u skupu S_1, tj. $(S_1 \subset S) \wedge (S_1 \neq S)$. Ukoliko nije reč o pravom podskupu nekog skupa, relacija inkluzije ili sadržavanja označava se sa $\supseteq$, tj. $\subseteq$. Npr. $A \supseteq B$ ili, što je isto, $B \subseteq A$. E je **pravi** podskup skupa P **ako** ("ako i samo ako") $E \supseteq P$ ali nije $P \supseteq E$.
-Prazan skup $\varnothing$ je podskup svakog skupa

Populacija mora biti definisana:

1. iscrpno (ne sme se izostaviti jedinka koja pripada populaciji ili dodati neka koja u nju ne pripada)

2. nedvosmisleno i

3. koncizno

Uzorak

Uzorak je deo populacije, podskup koji je izabran iz populacije i na kojem se vrši istraživanje. Uzorak je uređeni ili neuređeni skup entiteta iz populacije. Uređeni skupovi su jednaki ako sadrže iste članove koji su u istom redosledu. U psihologiji je najčešće uzorkovanje bez vraćanja, a ne sa vraćanjem.[9]

Slučajni uzorak je onaj koji se bira **slučajnim uzorkovanjem** (engl. random sampling). Slučajno uzorkovanje odvija se tako da svaki uzorak određene veličine (uređeni skup od n elemenata) iz neke populacije ima istu šansu da bude izabran. Definicija da je slučajni uzorak takav da "svaki elemenat populacije ima jednaku verovatnoću da bude izabran" tačna je samo kod uzorkovanja sa vraćanjem.[10]

Ako se iz populacije od N članova vrši slučajni izbor bez vraćanja uzorak je n-kombinacija elemenata populacije, a ako se vrši sa vraćanjem rezultat izbora je n-varijacija elemenata populacije.[11] Svaka n-varijacija elemenata populacije može biti izabrana sa verovatnoćom N^{-n} ili, što je isto, $1/N^n$. To je ekvivalentno sa uslovom da se u svakom pojedinačnom izboru element bira sa verovatnoćom N^{-1} pri čemu su pojedinačni izbori nezavisni. Rezultat svakog pojedinačnog izbora opisuje se slučajnom varijablom koja ima istu raspodelu kao i obeležje X.[12]

---

[9] Uzorkovanje može biti **bez vraćanja** i **sa vraćanjem**. Kod uzorkovanja sa vraćanjem član populacije se vraća u spisak iz kojeg se bira pre izvlačenja sledeće jedinice uzorka dok se kod uzorkovanja bez vraćanja član populacije koji je jednom izabran pri sledećem izvlačenju, ako bude opet izabran, preskače.

[10] Ako je izbor iz beskonačne (ili konačne) populacije učinjen sa vraćanjem, verovatnoća da neki uzorak od n jedinica bude izabran je n*(1/N) gde je 1/N verovatnoća za bilo kog člana populacije (veličine N) da bude izabran. Ukoliko je izbor učinjen iz konačne populacije bez vraćanja onda je verovatnoća izbora određenog člana u prvom izvlačenju 1/N, u drugom 1/(N-1), a pod uslovom da nije izvučen u prvom izvlačenju je složena verovatnoća (N-1/N) (1/N-1) = 1/N. Verovatnoća izbora u trećem izvlačenju je 1/(N-2), ali pod uslovom da nije izvučen u prvom i drugom izvlačenju ova verovatnoća je [1/(N-2)]* [(N-1)/N] *[(N-2)/N-1] = 1/N. Dakle, verovatnoća izbora odeređenog elementa ostaje ista (1/N) za svako izvlačenje i kod izvlačenja iz konačne populacije bez vraćanja, ali rezultati različitih izvlačenja nisu nezavisni.

[11] O pojmovima permutacija, varijacija i kombinacija može se pogledati u **Dodatku 2** ovog Priručnika. Inače, razumevanje daljeg teksta podrazumeva da je sadržaj Dodatka 2 poznat čitaocu.

[12] O slučajnim varijablama, funkcijama gustine i funkcijama distribucije može se pročitati u **Dodatku 2** ovog Priručnika.

Uzorak se može shvatiti kao n-torka (X_1, X_2, ..., X_n) gde su X_1, X_2, ..., X_n nezavisne slučajne varijable sa istom raspodelom kao i obeležje X.[13] Ova n-torka je prost slučajni uzorak. Realizovana vrednost uzorka (x_1, x_2,..., x_n) su konkretni brojevi i nisu slučajne varijable.

> **Zapamtite:** Treba razlikovati **slučajno uzorkovanje (random sampling)** od **slučajnog raspoređivanja (randomizacije,** engl. **random assignment** ili **randomization).** Slučajno raspoređivanje ili randomizacija je postupak nasumičnog, slučajnog raspoređivanja ispitanika po nivoima eksperimentalnog tretmana. Npr. od 90 ispitanika nasumice po 30 svrstamo u svaku od tri grupe, a svaka grupa dobija različit eksperimentalni tretman.

Jednostavni slučajni uzorak je onaj pri čijem izboru svaki uzorak veličine n ima istu verovatnoću izbora. Verovatnoća izbora svakog uzorka je $\frac{1}{\binom{N}{n}}$, jer je jednostavni slučajni uzorak kombinacija n elemenata od ukupno N.
Prema sličnom određenju jednostavni slučajni uzorak je onaj za kojega je verovatnoća za svakog člana populacije da bude izabran u uzorak ista [p ($e_i \in E$) = p $\forall$ $e_i \in P$].

Reprezentativni uzorak je deo populacije koji poseduje sve bitne osobine populacije. Da bi uzorak bio reprezentativan trebalo bi da u principu ne bude suviše mali (veći od 100).

Principi izbora reprezentativnog uzorka:

1. Svaki član populacije ima jednaku mogućnost da bude izabran -mora postojati tačan registar svih članova populacije.
2. Nepristrasnost pri izboru- izbor ne sme da zavisi od intervencije istraživača i ne sme izbor nekog člana zavisiti od izbora nekog drugog člana.

Metode slučajnog izbora: iz šešira, kocka, bubanj, tablica slučajnih brojeva, računar.
Metoda sistematskog izbora- bira se svaki n-ti član a početni član se bira slučajno.

Nereprezentativni uzorci:

Pristrasni- "telefonski";
Namerni- biranje za određene profesije (vojsku, policiju i sl.);
Prigodni- uzet s ruke -npr. brucoši za eksploraciju;

---

[13] Jednu od preciznih, ali početnicima teško razumljivih, definicija slučajnog uzorka dao je Viks (Wilks, 1962, str. 195): Ako je X jednodimenzionalna slučajna varijabla sa kumulativnom funkcijom gustine, tj. funkcijom distribucije F(x) u R_1, pri čemu je R_1 prostor mogućih ishoda, slučajni uzorak veličine n iz populacije sa funkcijom distribucije F(x) je n-dimenzionalna slučajna varijabla (X_1, X_2, ... , X_v) sa funkcijom distribucije $\prod_{i=1}^{n} F(x_i)$ u prostoru ishoda $R_n = R_1^{(1)} \times ... \times R_1^{(n)}$, pri čemu je $R_1^{(i)}$ jednodimenzionalni prostor ishoda, tj. realna brojna linija, za x_i, i = 1, ... ,n. (Oznaka $\prod_{i=1}^{n}$ označava proizvod elemenata x_i za raspon indeksa od 1 do n, a čita se kao "produkt *naziv opšteg člana* za i od 1 do n").

Na nereprezentativnim uzorcima smemo primeniti samo deskriptivnu statistiku.
Da bismo iz uzorka saznali o populaciji sa dovoljno malom greškom potrebno je da izbor članova populacije koji će ući u uzorak bude slučajan.

Varijabla v

Varijabla je karakteristika entiteta po kojoj se oni razlikuju. Ako svi entiteti nekog skupa imaju jednaku vrednost na nekoj "varijabli" onda je reč o **konstanti**. Određena vrednost koju ima neki entitet na nekoj varijabli je **vrednost varijable**.

Skup varijabli: $V = \{ v_j; j = 1,..., m\} \subset U$.

Uzorak varijabli biramo tako da bude reprezenatativan za univerzum varijabli (U) u skladu sa eksplicitno, jasno formulisanom teorijom.
Statistika se bavi postupcima pomoću kojih na nekom uzorku entiteta i uzorku varijabli utvrđujemo neku zakonitost i, ako je moguće, proveravamo pretpostavke zbog kojih smo počeli neku statističku analizu.

Podatak x_{ij}

$x_{ij} = e_i \otimes v_j$; $i = 1,...,n$; $j = 1,...,m$ (entitet e_i opisan na varijabli v_j).
Podatak je, dakle, određenje kategorijske pripadnosti na nominalnoj varijabli ili vrednost koju dobija ispitanik na nekoj kvantitativnoj varijabli.

Strukture podataka

1. Skup entiteta E opisan na jednoj varijabli v daje niz podataka u jednoj koloni[14]
$\mathbf{x} = E \otimes v =$

1	x_1
2	x_2
.	.
.	.
i	x_i
.	.
.	.
n	x_n

Na taj način se dobija struktura, tj. vektor varijable ili vektor podataka koji je reda nx1.[15]

---

[14] Opis skupa objekata na jednoj ili više varijabli je u stvari Dekartov proizvod dva skupa.
Dekartov proizvod:
Ako su A i B dva neprazna skupa onda se skup svih uređenih parova (a, b), gde je $a \in A$ i $b \in B$ zove Dekartov ili direktni proizvod skupova A i B: $A \otimes B = \{(a, b) \mid a \in A \wedge b \in B\}$.
Ako skup A sadrži n elemenata, a B m elemenata, Dekartov proizvod sadrži n x m elemenata.
[15] Red vektora ili matrice jednak je broju redova x broj kolona. Razumevanje struktura podataka i rada sa takvim strukturama podrazumeva da su, iz bilo kojeg štiva, čitaocu poznati osnovi linearne algebre.

2. Uzorak entiteta E opisan u pogledu više varijabli

$\mathbf{X} = E \otimes V =$

	1	2	.	.	j	.	m
1	x_{11}	x_{12}			x_{1j}		x_{1m}
2	x_{21}	x_{22}					x_{2m}
.							.
.							.
i	x_{i1}	.			x_{ij}		x_{im}
.							.
.							.
n	x_{n1}				x_{nj}		x_{nm}

Na taj način se dobija **X**, matrica podataka reda n x m. Očigledno je da je prvi slučaj specijalni slučaj ovog opštijeg kada je m =1.
Opis uzorka u pogledu jedne ili više varijabli zapravo znači **merenje** svakog člana uzorka u pogledu neke varijable.

Merenje

Kembel: **Merenje** je pridavanje brojki pojavama ili svojstvima objekata na osnovu određenih pravila[16].

Momirović, Wolf i Popović : Merenje je (prema najširoj definiciji) "...svaka operacija, koja , u skladu sa nekim potpunim i tačnim skupom pravila, omogućava da se nekom objektu koji je član nekog homogenog skupa objekata pridruži oznaka ili broj koji se odnosi na određeno svojstvo, tako da se ma koja dva objekta koja se po tom svojstvu razlikuju mogu razlikovati, i ma koja dva objekta koja su po tom svojstvu identična mogu smatrati identičnim." (Momirović, Wolf & Popović, 1997, str.19)
"Merenje se može shvatiti kao operacija preslikavanja nekog objekta iz skupa objekata E na skup realnih brojeva , učinjena u skladu sa nekom procedurom P." (Momirović et al., str.20).
U užem smislu, u smislu merenja kvantitativnih svojstava Momirović et al. formalno definišu merenje na sledeći način:
"Neka je e_i neki objekt iz skupa objekata $E = (e ; i = 1,...,n)$ i neka je w neka karakteristika definisana na skupu realnih brojeva $(x_j , \infty < x_j < \infty)$. "Varijabla koja je "definisana pridruživanjem broja x_j objektu e_i

$$y_{ij} = e_i \otimes x_j$$

" je " rezultat merenja karakteristike w na objektu e_i (isto, str.20). Merimo osobine, karakteristike, obeležja (to su predmeti merenja) članova populacije koji su ušli u uzorak i dobijamo određene podatke o ovim obeležjima. Podaci mogu biti kvalitatitivni i kvantitativni.
Merni instrument je ono čime merimo bilo da je to aparat (npr. reakciometar), upitnik i sl.

---

[16] Ne postoji potpuno slaganje o tome šta sve ima pravo da se zove merenjem u psihologiji. Pojedini autori priznaju kao merenje samo merenje **kvantiteta**. Aksiomatska definicija kvantiteta i za psihologe veoma korisna rasprava o problemu merenja može se pročitati u: Michel, J. (1990).*An introduction to the logic of psychological measurement*, Hillsdale, NJ:Erbaum.

Merenjem dobijamo dve vrste mera: mere date u kontinuiranom nizu i mere date u diskontinuiranom (diskretnom) nizu mera.
Kontinuirani niz mera potiče od osobine koja se može predstaviti duž kontinuuma, tj. dimenzije na kojoj svaka tačka predstavlja realnu meru predmeta merenja. Sledeća obeležja daju kontinuirani niz mera: prostor, vreme, kognitivne funkcije i većina osobina ličnosti. Grafički prikazano to izgleda ovako

---

Kod kontinuiranih varijabli između dve vrednosti varijable može biti neodređeno mnogo vrednosti. Mera je kod kontinuiranih varijabli stoga predstavnik celog intervala između gornje i donje egzaktne granice.
Diskretni nizovi mera potiču od predmeta merenja koji mogu imati realne vrednosti samo u pojedinim tačkama brojne linije: broj dece u porodici, broj grla stoke u domaćinstvu, broj bodova sa utakmice. Grafički se diskretni nizovi mera mogu predstaviti ovako: . . . .

Vrste numeričkih podataka

Postoje dve vrste numeričkih podataka:

1. **frekvenca** f (često se naziva i učestalošću i obeležava slovom **u**).
2. **mere ili skorovi** (x) koji se dobijaju sa merne skale nekog mernog instrumenta.

Merna skala mernog instrumenta je standardizovani numerički sistem pomoću kojeg se ostvaruje merenje.

Merne skale

Momirović, Wolf i Popović dele merne skale na:

1.Kardinalne (apsolutne). Na njima je definisana operacija prebrojavanja i relacija identiteta.[17] Svojstva uređenosti i intervalnosti zadržavaju se samo pri identičnim transformacijama $x_{ij}= y_{ij}$. Ove skale imaju realnu nultu tačku, tj. vredi

$$y_{ij} = 0 \Rightarrow v_j=0$$

2.Aditivne skale zadržavaju uređenost i intervalnost pri dodavanju i oduzimanju, tj. pri transformaciji tipa: $y_{ij}= x_{ij} - c$, gde je c proizvoljna nenulta realna konstanta. Nemaju realne nule pa množenje konstantom nema smisla.

3.Intervalne skale nemaju realne nule a merna jedinica je arbitrarna. Zadržavaju svojstva uređenosti i intervalnosti pri transformacijama tipa $y_{ij} = (x_{ij} - c)\, m^{-1}$, gde je c proizvoljna realna konstanta, a m nenulta realna konstanta.

4.Razmerne skale imaju realnu nulu, merne jedinice su im definisane na kontinuumu a jedinice konvencionalno zavise od načina merenja. Dozvoljeno je množenje i deljenje bilo kojom realnom konstantnom različitom od 0. Dodavanjem i oduzimanjem skala prestaje da bude razmerna.

---

[17] O relacijama i operacijama može se pročitati u B. Boričić i M. Ivović (1999). *Matematika*, Ekonomski fakultet, Beograd.

Nivoi merenja

Ponekad se, u vezi sa mernim skalama, govori o nivoima merenja. Čak i laiku je jasno da nije isto značenje cifara u sledećim primerima:

Nacionalna pripadnost:
1.Englez
2.Šveđanin
3. Srbin
4. Grk...

Stručna sprema
1.NSS
2.SSS
3.VŠS
4.VSS

Uspeh na Testu znanja:
1 bod
2 boda
8 bodova
43 boda...

Visina
167 cm
187 cm
188 cm

Svakom nivou merenja odgovara određena skala merenja:
Prema Stivensu "skala merenja je pravilo po kojem se brojevi pridružuju određenim aspektima objekata ili događaja" (prema Stilson, 1966., str.126).
Nivoi merenja zavise od prirode obeležja koje merimo, tj. varijable.
Često se razlikuju 4 nivoa merenja sa različitim skalama merenja i sa razlikama u količini informacija koje merilac dobija primenom svake od skala. Svaki tip skale definisan je svojim invarijantama (prema Kornbrot, 1990).

I.**Nominalna** skala (lat. Nomen = ime; nominalis = imenski) je najniži tip skale. Cifra ovde služi samo kao ime kategorija datog obeležja. Cifre nemaju kvantitativno značenje već govore samo da je jedno isto ili različito od drugog: $A = A$ (identitet) i $A \neq B$ (različit od).
Nema relacije $A > B$ i $A < B$. Dakle, jedina relacija je ekvivalencija koja je refleksivna ($X = X$ za svako X) , simetrična ($X=Y \Rightarrow Y=X$) i tranzitivna (Ako je $X=Y$ i $Y=Z$ onda je $X=Z$).
Ova skala invarijantna je na permutaciju svojih vrednosti. Dakle, dozvoljena je operacija međusobne zamene oznaka kategorija ako je zamena izvedena iscrpno i konzistentno.

II. **Ordinalna** (lat. Ordo = red; ordinalis = redni) skala daje informaciju o redosledu. Nepoznata je veličina njenih jedinica. Nemaju svojstvo intervalnosti ni realnu 0. Za ordinalne skale važi relacija potpunog ili linearnog poretka ($\geq$ ili $\leq$) koja je tranzitivna, antisimetrična (Ako je $X \geq Y$ i $Y \geq X$, onda je $X = Y$) i linearna ($X \geq Y$ ili $Y \geq X$). Ove skale zadržavaju svojstvo uređenosti nakon transformacija tipa

$$y_{ij} = f(x_{ij})$$

gde je f monotona (rastuća ili opadajuća) funkcija. Skala je invarijantna na monotone transformacije (koje zadržavaju redosled) kao npr.. logaritamske, ekponencijalne ili inverzne. Pojedine crte ličnosti kao što su poštenje, hrabrost i sl. sada se mere na ordinalnom nivou, tj.skalama procene.

III. **Intervalna** skala (lat. Intervallum = prostor između bedema, međuprostor)
Njene jedinice su ekvivalentne, veličina jedinica i nulta tačka arbitrarne. Ova skala mora imati (pored relacije ekvivalencije i potpunog poretka postoji specifikovan racio bilo koja dva intervala. Daje informaciju o tome šta je veće ili manje i za koliko jednakih jedinica je nešto veće ili manje. Invarijantna je na linearne transformacije tipa

$$y_{ij} = f(x_{ij}) = a + bx_{ij}.$$

Primer: temperatura u F^o. Porast u prosečnoj molekularnoj kinetičkoj energiji koji je neophodan da podigne temperaturu tela sa 1 na 2 stepena isti je kao i onaj koji je neophodan za dizanje temperature sa 18 na 19 stepeni.
Za intervalne varijable može se formirati racio skala razlika između vrednosti varijable. Razlike prave racio skalu jer oduzimanje poništava arbitrarnu nultu tačku (prema Ferguson, str.14). Razlike su iste bez obzira na početnu tačku.

IV. **Racio (razmerna)** skala (lat. ratio = količnik). Ima ekvivalentne i ekvidistantne jedinice i realnu apsolutnu nulu dakle vredi $y_{ij} = 0 \Rightarrow v_j = 0$
Daje informacije i za koliko puta je nešto veće ili manje. Invarijantna je na multiplikativne transformacije tipa

$$y_{ij} = f(x_{ij}) = mx_{ij}$$

pri čemu je m proizvoljna realna nenulta konstanta.

Parametar i statistik

Parametar je statistička mera numeričke karakteristike populacije. Statistička mera numeričke karakteristike uzorka je **statistik**. Statistici uzorka su **ocenjivači (estimatori)** parametara, a konkretne vrednosti statistika koje su dobijene na nekom uzorku su **ocene** parametra.[18]

Stepeni slobode (df)

Da bi se izračunao neki statistik potrebno je koristiti opservacije dobijene na uzorku i određene parametre. Ako ne znamo parametre onda se oni ocenjuju na osnovu uzorka. Broj stepeni slobode (df) je broj nezavisnih opservacija u uzorku umanjen za broj parametara koji se moraju oceniti na osnovu uzorka.

[18] O ovome će biti podrobnije reči u glavi VI kada bude govora o statističkom ocenjivanju parametara.

IV. SREĐIVANJE PODATAKA - STATISTIČKI OPIS (I GRAFIČKI PRIKAZ) POPULACIJE ILI UZORKA U POGLEDU JEDNE KVANTITATIVNE VARIJABLE

Struktura podataka:

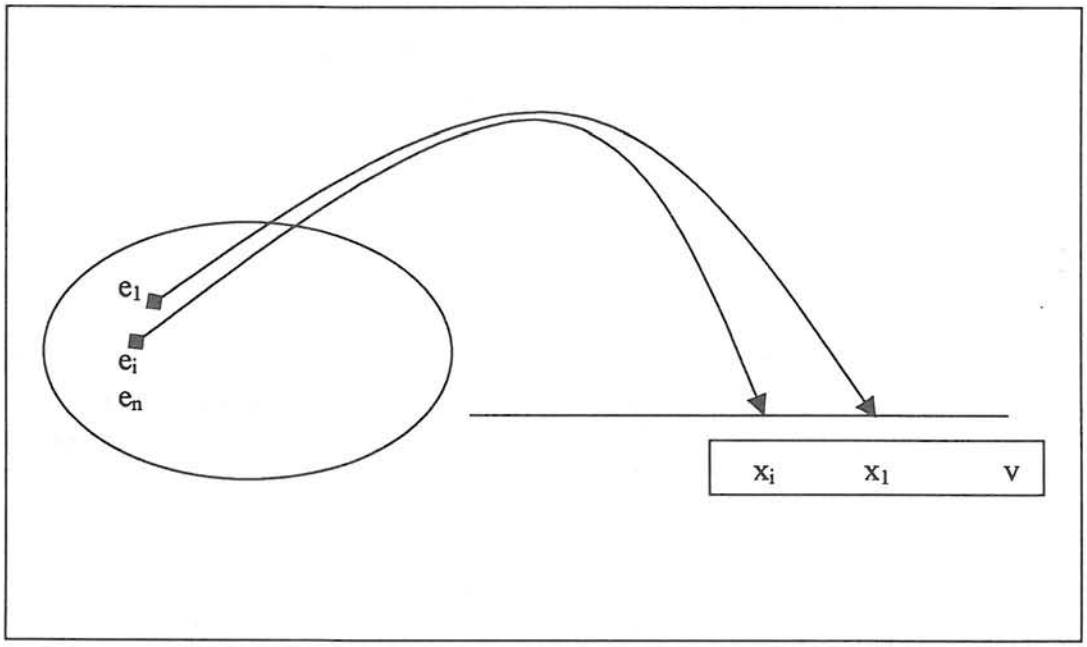
Dati su jednostavni slučajni uzorak objekata [$E = \{ e_i ; i = 1,...,n\} \subset P$] i varijabla v kvantitativnog karaktera na kojoj rezultati označavaju veličinu ove karakteristike za dati entitet.

Opisujemo skup E na varijabli v. Ovaj opis je, u stvari, **preslikavanje** elemenata skupa E u skup vrednosti varijable v.[19] Ishod ovog preslikavanja je vektor $\mathbf{x}$, čiji elementi su **rezultati** ili **mere** (x_i) ispitanika (slika 5).

$E \otimes v = \mathbf{x}$

$\mathbf{x} = (x_i); i = 1,2,\ldots, n$

Slika 5: Preslikavanje elemenata uzorka E na skup vrednosti varijable v



Rezultate ispitanika na varijabli v možemo tretirati kao elemente kolonskog vektora.

---

[19] Ako su E i X skupovi, relacija f iz E u X je podskup Dekartovog proizvoda skupova E i X, j. $f \subset E \otimes X$. Relacija f je **funkcija** ili **preslikavanje** ako se po određenom pravilu svakom elementu skupa E pridružuje određeni element skupa X. Preciznije, relacija f je funkcija iz E u X (najčešće se piše kao $f: E \rightarrow X$) ako važi sledeće:

1. $(\forall e \in E)\ (\exists x \in X)\ (e, x) \subset f$. Dakle, za svaki član e iz skupa E postoji x iz skupa X takvo da e i x daju uređeni par koji je podskup relacije f;
2. $(e, x_1) \in f \wedge (e, x_2) \in f \Rightarrow x_1 = x_2$. Dakle, jednom elementu skupa E ili **domena** funkcije f odgovara jedan i samo jedan element skupa X, ili **kodomena** funkcije f.

IV.1.Uređivanje podataka - sortiranje: uzlazno (a) i nizlazno (d)

****

U SPSS programu sortiranje se vrši na sledeći način: U meniju **Data** odabrati **Sort Cases.** U meniju za dijalog **Sort Cases** u polje **Sort by** prebaciti (bilo dvostrukim klikom na njeno ime u levom okviru, bilo jednim klikom na ime varijable, a potom na strelicu između levog i desnog okvira) iz levog spiska ime varijable po kojoj treba sortirati podatke, a u okviru **Sort order** odabrati (klikom na dugme koje treba da se zatamni) **Ascending** ili **Descending** opciju. Ascending (uzlazna) opcija će ispitanika sa najnižim rezultatom staviti na prvo mesto, a descending (nizlazna) opcija će kao prvi rezultat staviti najviši rezultat.

****

IV.2.Klasifikovanje podataka - pravljenje raspodele učestalosti

Da bi se bolje razumele informacije sadržane u skupu podataka dobijenih merenjem nekog svojstva na uzorku entiteta često je potrebno te rezultate urediti, klasifikovati i sažeto statistički opisati.

Utvrđivanje najmanjeg (x_{min}) i najvećeg (x_{max}) dobijenog rezultata

$$x_{min} = \min_i x_i, \; i = 1, \ldots, n$$

$$x_{max} = \max_i x_i, \; i = 1, \ldots, n$$

Koliki je najveći, a koliki najmanji rezultat moguće je jednostavno videti iz sortiranih podataka na varijabli.

Pravljenje raspodele učestalosti

Raspodela učestalosti predstavlja takvo uređenje podataka koje pokazuje učestalost tj. **frekvencu** (f_k) pojedinih vrednosti (ili vrednosti unutar intervala) varijable.
Ukoliko se ne želi jedinična raspodela (prebrojavanje učestalosti svake vrednosti ponaosob) potrebno je odrediti veličinu razreda i broj razrednih intervala.[20]
Izgled distribucije frekvenci (raspodele učestalosti) kada su vrednosti na varijabli celobrojne prikazan je u donjoj tabeli. Oznaka i u toj tabeli je veličina razrednog intervala, k je oznaka k-og razreda, a l je oznaka najvišeg razreda.

Razredi	Frekvence razreda
$(x_{min}, x_{min} + i - 1)$	f_1
$(x_{min} + i, \; x_{min} + 2i - 1)$	f_2
$(x_{min} + i, \; x_{min} + ki - 1)$	f_k
.	
$(x_{min} + (l-1)i, \; x_{min} + li - 1)$	f_l

Relativna frekvenca: $p_k = f_k / n$

[20] Praktična uputstva o ovome se mogu naći u knjizi Statistika za psihologe Č. Dragićevića.

Ako je uzorak zaista slučajan relativna frekvenca je ocena verovatnoće da neki rezultat padne u neki razred.[21]

Kumulativna frekvenca: $cf_k = \sum_{j=1}^{k-1} f_j + f_k = \sum_{j=1}^{k} f_j$

$cf_1 = f_1$

$cf_2 = f_1 + f_2$

$cf_3 = f_1 + f_2 + f_3$

.....

$cf_l = f_1 + f_2 + \cdots + f_k + \cdots + f_l$

Očigledno, kumulativna frekvenca u najvišem razredu (k = l) jednaka je broju rezultata, tj. $cf_l = n$.

Relativna kumulativna frekvenca - $rcf_k = cf_k / n$.

****

Za pravljenje raspodele učestalosti u SPSS programu neophodno je u meniju **Statistics** odabrati podmeni **Summarize**, a potom u podmeniju **Frequencies** odabrati opciju **Display frequency tables**. Ova opcija je obično automatski uključena tj. ispred nje stoji kvačica na Windows95 ili putača na Windows3.xx sistemima.

Na ovaj način program će napraviti jediničnu raspodelu učestalosti tako što će prebrojati koliko se često svaki rezultat pojavljuje. Ukoliko ima veliki broj različitih vrednosti onda je potrebno varijablu za koji se želi raspodela koja se sažima u manji broj razreda prethodno rekodirati prema uputstvima koja će biti data u odeljku **IV.3.** (meni **Transform** i podmeni **Recode**) i potom dati Labele tj. nazive razreda -opcija **Value labels** u meniju **Data** - podmeni **Define Variables**. Doduše, u meniju **Statistics/Summarize/Explore** moguće je klikom na dugme **Statistics** otići u poddijalog **Explore: Statistics** i tamo odabrati opciju **Grouped frequency tables** i na taj način zatražiti da program sam sažme mere u razrede i napravi distribuciju učestalosti sa tako dobijenim razredima. Međutim, retko kada ćete biti zadovoljni načinom na koji je grupisanje mera u razrede izvršeno, tako da je ovu mogućnost bolje i ne uzimati u obzir.

[21] Pojedini autori, frekvencu označavaju slovom u (učestalost), a relativnu frekvencu slovom f.

Primer: Kao ishod procedure **Frequencies**, distribucija varijable sa IMENOM (name) EKSTRA, čiji je NAZIV (label) AJZENK- EKSTRAVERZIJA izgleda ovako:

```
EXTRA      AJZENK_EKSTRAVERZIJA
```

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	2	2	.6	.7	.7
	3	1	.3	.3	1.0
	4	4	1.3	1.3	2.3
	5	7	2.2	2.3	4.6
	6	6	1.9	2.0	6.6
	7	11	3.5	3.6	10.3
	8	8	2.5	2.6	12.9
	9	12	3.8	4.0	16.9
	10	14	4.4	4.6	21.5
	11	20	6.3	6.6	28.1
	12	23	7.3	7.6	35.8
	13	25	7.9	8.3	44.0
	14	23	7.3	7.6	51.7
	15	30	9.5	9.9	61.6
	16	25	7.9	8.3	69.9
	17	35	11.1	11.6	81.5
	18	27	8.5	8.9	90.4
	19	17	5.4	5.6	96.0
	20	12	3.8	4.0	100.0
	.	14	4.4	Missing	
	Total	316	100.0	100.0	

```
Valid cases      302      Missing cases      14
```

U koloni **Value** su mere koje su ispitanici postigli na ekstroverziji merenoj Ajzenkovim upitnikom EPQ, tj. vrednosti varijable EKSTRA koje se pojavljuju u podacima. U koloni **Frequency** vidimo koliko ispitanika iz uzorka ima datu meru ekstroverzije, a u koloni **Percent** relativnu frekvencu mere (u odnosu na ukupno n = 316). Relativna frekvenca date mere u odnosu na ukupan broj ispitanika za koje imamo mere na toj varijabli (**Valid cases** = n - **Missing cases** = 302) data je u koloni **Valid percent** a kumulativna relativna frekvenca (u odnosu na broj ispitanika za koje **imamo** podatke) nalazi se u koloni **Cum Percent**.

****

IV.3. Transformisanje unetih vrednosti varijable

Rekodiranje vrednosti varijable-komanda RECODE

Ponekad je neophodno pretvoriti pojedine vrednosti (ili čitav raspon vrednosti) neke kvantitativne varijable u novu vrednost ili, ako je reč o kategoričkoj varijabli, zameniti cifarske oznake kategorija nekom drugom cifrom. Tako, na primer u slučaju pravljenja raspodele učestalosti koja ima intervale razreda neophodno je, kako smo već istakli, raspon vrednosti koje ulaze u isti interval rekodirati u jedan broj. Isto tako, pri unosu podataka sa

nekog upitnika najlakše nam je da različite odgovore zamenimo pri unosu ciframa od 1 pa nadalje, idući sleva udesno. Kada, pak, hoćemo da sabiranjem vrednosti odgovora na svaku stavku, u skladu sa ključem datog upitnika, izračunamo ukupni rezultat ili skor, ponekad moramo unete vrednosti pre računanja skora da rekodiramo. Ako smo, na primer, odgovor DA na svakoj stavki kodirali i unosili kao cifru 1, a odgovor NE kao cifru 2 onda ćemo za neke stavke, pre računanja ukupnog skora na upitniku, oznake 2 ili 1 na pojedinim stavkama pretvoriti u nule, a na drugim stavkama u jedinice. To ćemo uraditi u skladu sa ključem za bodovanje odgovora DA i NE na stavkama upitnika.

****

Rekodiranje varijable u SPSS programu vrši se odabirom **Transform/Recode**.
Ukoliko odaberemo opciju **Into Same Variables** stare vrednosti u **postojećoj** varijabli biće zamenjene novim. Pri odabiru ove opcije pojavljuje se dijalog prozor **Recode into Same Variables**. U okviru uz levu ivicu prozora je dat spisak sa imenima (**name** a ne **label**!) svih varijabli koje se nalaze u fajlu sa podacima. Imena varijabli koje se rekodiraju treba prebaciti iz ovog spiska u desni okvir (bilo dvostrukim klikom na ime varijable u levom okviru, bilo jednim klikom na ime varijable u levom okviru pa potom na dugme-strelicu između levog i desnog okvira). Moguće je klikom na dugme **If** definisati uslov koji treba da ispuni ispitanik da bi se vrednost varijable za datog ispitanika promenila. Ukoliko se vrši uslovno rekodiranje u dijalog prozoru koji se dobije klikom na dugme **If** (naslov ovog prozora je **Recode into Same Variables: If case**) treba kliknuti na dugme ispred opcije **Include if case satisfies condition:**, a zatim u praznom okviru ispod nje uneti uslov. Naprimer, ako se postavlja uslov da se vrednosti neke varijable rekodiraju samo za muskarce treba upisati **pol =1**. Podrazumeva se da su u varijabli pol muškarci označeni jedinicom. Posle postavljanja uslova, na osnovni (prethodni) dijalog za rekodiranje vraćamo se klikom na dugme **Continue**. Ukoliko ne želimo da se izvršava postavljeni uslov možemo kliknuti na dugme **Cancel** da bismo se vratili na osnovni dijalog za rekodiranje.
Definisanje načina rekodiranja vrednosti na nekoj varijabli (bilo da je rekodiranje uslovno ili važi za sve objekte posmatranja) vršimo tako što klikom na dugme **Old and New Values** (dijaloga **Recode into Same Variables**) pređemo u poddijalog **Recode into Same Variables:Old and New Values** i definišemo željene transformacije. U levom okviru (Old Value) odaberemo (zavisno od potrebe) jednu od opcija (**Value, System missing, System or user missing, Range, All other values**), a u okviru u desnom gornjem uglu (**New value**) odabiramo ili **Value** ili **System missing** opciju. Posle svake definicije transformisanja jedne (ili više) starih vrednosti u jednu novu (naprimer da stara vrednost 5 postaje nova vrednost 1) treba kliknuti na dugme **Add** i definicija rekodiranja će biti upisana u desni okvir **Old → New:**. Ukoliko neku od definicija rekodiranja koje su već upisane u okvir **Old → New:** želimo da promenimo treba je odabrati (jednim klikom miša na nju), napraviti ispravku u poljima okvira Old value i New value i zatim kliknuti na dugme **Change**. Za brisanje već upisane definicije iz okvira treba (po njenom odabiranju) kliknuti na dugme **Remove**. Za vrednosti koje ostaju iste nije potrebno ništa definisati.
Po završenom definisanju svih željenih transformacija klikom na dugme **Continue** vraćamo se iz poddijaloga u glavni dijalog **Recode into Same Variables** i klikom na dugme **OK** (ili pritiskom na taster **Enter** na tastaturi) završavamo definisanje operacije i dajemo SPSS-ovom procesoru do znanja da treba da izvrši komandu.
Ukoliko želimo da zatvorimo prozor za dijalog ali tako da se definisane transformacije **ne** izvrše treba kliknuti na dugme **Cancel** (ili **Esc** na tastaturi). Ako, pak, želimo da dijalog

prozor bude i dalje otvoren, a da poništimo date definicije rekodiranja treba kliknuti na dugme **Reset**.

> **Zapamtite:** Klikom na dugme **Paste** u bilo kojem prozoru za dijalog SPSS-a, u kojem ono postoji, možemo dobiti tekst do tog trenutka definisane komande u sintaksnom prozoru. Tako dobijene komande moguće je sačuvati u nekom fajlu i kasnije ih (modifikujući ih po potrebi) izvršavati u sintaksnom načinu rada.

> **Zapamtite:** Funkcije dugmadi **If, Paste, Cancel, Reset, Continue, OK, Help, Add Change, Remove** i **Help** su suštinski iste u svim prozorima za dijalog u kojima s pojavljuju.

> **Zapamtite**: Iz jednog u drugo ulazno polje unutar određenog dijalog prozora možemo ići pritiskom na dugme **Tab** na tastaturi. U prethodno polje se možemo vratiti istovremenim pritiskom na dugmad **Shift** i **Tab**.

U slučaju da želimo da sačuvamo u fajlu sa podacima nerekodiranu varijablu ali i da od nje napravimo novu sa rekodiranim vrednostima stare varijable odabraćemo u meniju **Transform/ Recode** opciju **Into Different Variables**. U okvir **Input Variable→Output Variable:** prebacićemo (iz spiska u levom okviru) ime izvorne (ulazne) varijable, a u polju **Name** (u okviru **Output Variable**) upisaćemo ime nove varijable (ukoliko želimo njen širi naziv upisaćemo ga u polju **Label**) i na kraju kliknuti na dugme **Change**. U tom trenutku će u okviru **Input Variable→Output Variable:** znak pitanja posle strelice biti zamenjen imenom nove (izlazne) varijable. Definisanje uslova (poddijalog **Recode Into Different Variables: If Cases**) i načina rekodiranja (poddijalog **Recode into Different Variables:Old and New Values**) je praktično isto kao kod rekodiranja unutar iste varijable. Bitna razlika je jedino u tome što sudbina **svih** vrednosti izvorne varijable mora biti definisana. Naime, za vrednosti koje ostaju iste u izlaznoj varijabli kakve su bile u izvornoj potrebno je (kada ih unesemo u polje okvira **Old Value**) u okviru **New Value** odabrati opciju **Copy old value(s)** i zatim kliknuti na **Add**. Ukoliko to ne bismo uradili vrednosti koje bi trebalo da budu nepromenjene u izlaznoj varijabli (u odnosu na izvornu) bile bi zamenjene sistemskim nedostajućim vrednostima **(System missing values)**.

****

Izvođenje nove varijable-komanda COMPUTE

Jedna od često korišćenih transformacija podataka sastoji se od izvođenja nove varijable na osnovu postojećih. Naprimer, vrlo često želimo da na osnovu unetih odgovora na pojedinim stavkama upitnika izračunamo ukupnu meru (ili skor) na dimenziji koju upitnik meri. Takođe, veoma često nam je potrebno da na osnovu nekoliko reakcija na dati stimulus (a pokazatelji pojedinačnih reakcija su upisani u zasebne varijable) izračunamo za svakog ispitanika njegovu prosečnu vrednost (i standardnu devijaciju) pojedinačnih reakcija i da tako dobijeni rezultat upišemo kao novu varijablu. Za ove i razne druge transformacije najbolje je koristiti komandu **Compute**.

****

U meniju SPSS programa biramo **Transform/Compute** a zatim, u dijalogu **Compute Variable**, u ulazno polje **Target variable** upisujemo ime nove, ciljne varijable. (Ako želimo, mada to nije obavezno, klikom na dugme **Type & Label** možemo otići u poddijalog **Compute Variable:Type & Label** u kojem možemo uneti širi naziv i definisati tip nove varijable). U ulaznom okviru **Numeric Expression** treba definisati transformaciju kojom će nastati ciljna varijabla. Definicija transformacije obuhvata unošenje (bilo iz spiska u levom okviru bilo upisivanjem sa tastature) imena varijabli čijom transformacijom nastaje nova i simbola matematičkih operacija koje treba nad njima načiniti. Simboli operacija mogu se upisivati sa tastature ili klikom miša na odgovarajuće dugme u skupu dugmadi ispod ovog okvira. Tako, ako bismo hteli da nova varijabla ima vrednosti koje su prosto 2.3 puta veće nego vrednosti na postojećoj varijabli LKL upisali bismo u polje **Numeric Expression** 2.3*LKL.
Moguće je (klikom na dugme **If**) postaviti (u poddijalogu **Compute Variable: If Cases**) uslov koji treba da ispune ispitanici za koje će se računati vrednosti na novoj varijabli.
Isto tako, u okviru **Functions:** na raspolaganju je veliki broj funkcija za transformaciju. Po odabiru željene funkcije iz spiska postojećih funkcija treba kliknuti na dugme sa strelicom nagore da bi se funkcija upisala u okvir **Numeric Expression**. Po unosu funkcije u okvir potrebno je (u skladu sa sintaksom date funkcije) znake pitanja u zagradi zameniti odgovarajućim argumentima (npr imenima varijabli).Naprimer, za računanje nove varijable **UKESKOR**, koja je prosti zbir varijabli E1 do E30 možemo (umesto da upišemo E1+E2+E3...) u okvir **Numeric Expression** uneti funkciju **SUM(?,?)** i zatim umesto znakova pitanja upisati E1,E2,...ili kraće E1 **TO** E30. Spisak i kraća objašnjenja pojedinih funkcija možemo dobiti klikom na dugme **Help** u okviru dijaloga **Compute Variable**.
****

Komanda COUNT

Ovom komandom možemo napraviti novu, ciljnu varijablu u kojoj će biti upisano koliko puta se neka vrednost pojavljuje kod nekog ispitanika u skupu **numeričkih** varijabli. Naprimer, možemo prebrojiti na koliko pitanja od ukupno 30 pitanja (a svako pitanje je uneto kao posebna varijabla) je svaki ispitanik odgovorio tačno. Ako su odgovori tačno unošeni kao jedinice prebrojili bismo koliko jedinica svaki ispitanik ima na svim pitanjima i ta vrednost bi bila upisana kao vrednost koju ispitanik ima na novoj varijabli.
****
Odabirom menija **Transform/Count** dobijamo dijaloški prozor **Count Occurences of Values within Cases** u kojem u ulazno polje **Target Variable** upisujemo ime nove, ciljne varijable (ako želimo širi naziv možemo ga upisati u polje **Target Label**). Zatim u okvir **Numeric Variables:** prebacimo spisak varijabli u kojima će se tražiti da li data vrednost postoji, a klikom na dugme **Define Values** pozivamo poddijalog **Count Values within Cases:Values to Count**. U ovom poddijalogu posle odabira vrednosti čiju pojavu treba registrovati u varijablama iz spiska treba kliknuti na dugme **Add**. Da bi program izvršio prebrojavanje i u novu varijablu upisao za svakog ispitanika učestalost pojavljivanja date vrednosti u skupu varijabli po završenom poslu u poddijalogu kliknemo na **Continue**, a u osnovnom dijalogu na **OK**.
****

Komanda Automatic Recode

Premda ovom komandom možemo vrednosti na bilo kakvoj varijabli pretvoriti u varijablu sa nizom celih brojeva, ona je najkorisnija u situaciji kada oznake na nekoj tekstualnoj (string) varijabli želimo da pretvorimo u cifre. Naprimer, ako smo pol definisali kao tekstualnu (string) varijablu i u njoj unosili podatke u obliku MUSKO i ZENSKO (mada je to stvarno glupa i besmislena opcija!) možemo, kada nam je to neophodno, ovom komandom unesene tekstualne oznake pretvoriti u cifre 1 i 2. Dakle, ova komanda različite vrednosti numeričkih i tekstualnih varijabli pretvara u cele brojeve.

****

Po odabiru menija **Transform/Automatic Recode** treba u prozoru za dijalog **Automatic Recode** u okviru **Variable New Name:** ubaciti ime varijable čije vrednosti želimo da rekodiramo na ovaj način, zatim u polje iza dugmeta **New Name** upisati ime varijable u koju će biti upisane nove vrednosti i kliknuti na dugme **New Name**. U okviru **Recode starting from** treba odabrati jednu od opcija. Ako je reč o tekstualnoj varijabli odabir jedne od ovih opcija je nevažno. Međutim, ako je varijabla koju rekodiramo numerička odabir opcije **Lowest value** znači da će najniža vrednost na varijabli biti pretvorena u jedinicu a opcija **Highest value** će dovesti do toga da najveća vrednost u staroj varijabli postane u novoj varijabli vrednost jedan.

Napomena: Kada izvrši automatsko rekodiranje program će u prozoru za ispis (Output window) dati obavešenje kako je izvršeno rekodiranje. Ovaj ispis je korisno zapisati u nekom fajlu da bismo pri tumačenju rezultata analiza u kojima je korišćena nova varijabla mogli da se snađemo.

****

IV.4. Mere centralne tendencije

Nizovi numeričkih podataka, dobijeni posmatranjem ili merenjem pojava redovno pokazuju tendenciju grupisanja oko neke centralne vrednosti. Statističke mere koje su izraz ove tendencije su mere proseka ili mere centralne tendencije. One služe sažetom opisivanju mnoštva podataka o nekoj pojavi.

Često se u svakodnevnom životu čuje izraz "prosečan". Pri tom se najčešće misli na aritmetičku sredinu. No to nije jedina mera proseka.

Izbor mere centralne tendencije zavisi od prirode varijable (vrste merne skale sa koje mere potiču) i oblika distribucije varijable.

Aritmetička sredina (M ili AS)

Uslovi za njeno korišćenje:

-mere moraju poticati bar sa intervalne merne skale;

-raspodela ne sme imati ekstremne rezultate u jednom smeru.

Računanje AS iz sirovih podataka:

$$M = \sum_{i=1}^{n} x_i * n^{-1}$$

Aritmetička sredina varijable u matričnom obliku:[22]

$$M = \mathbf{e}^t\mathbf{x} * n^{-1}$$

gde je **e** vektor sumiranja (vektor jedinica reda n), **x** vektor sa rezultatima za n ispitanika na varijabli X, a n veličina uzorka.
Osobine aritmetičke sredine:

- Njome se lako može manipulisati algebarski jer je definisana jednačinom;
- Stabilnija je od drugih mera centralne tendencije, tj. ima manju standardnu grešku;[23]
- Bolja je ocena populacijske AS nego Mod i Mdn;[24]
- AS je veća od najmanje i manja od najveće mere u nizu

$$x_{min} < M < x_{max}$$

- Dodavanje konstante svakoj meri iz niza povećava AS za iznos konstante:

$$M(x + a) = M(x) + a$$

- Množenje svake mere iz niza konstantom povećava AS za konstantu puta:

$$M(x * a) = M(x) * a$$

- Zbir odstupanja svih mera iz niza od njihove AS je 0. Prema tome AS je težište svih mera u nizu:

$$\sum_{i=1}^{n}(x_i - M_x) = 0$$

- Suma **kvadriranih** odstupanja svake mere od AS je manja nego od bilo koje druge vrednosti: AS je centroid ili centar gravitiranja niza mera.

$$\sum_{i=1}^{n}(x_i - M_x)^2 = \text{minimum} < \sum_{i=1}^{n}(x_i - x_0) \,|\, x_0 \neq M_x$$

---

[22] Za razumevanje matrične notacije, koja će biti često korišćena u ovom Priručniku, potrebno je savladati elemente linearne algebre. Elementi linearne algebre prikazani su u dodatku knjizi *Teorija merenja I* K.Momirovića, B. Wolfa i D. A Popovića. Vrlo korisna knjiga na engleskom jeziku za savladavanje ovih elemenata je: Green, P.E & Carroll, J.D. (1976). *Mathematical tools for applied multivariate analysis*, London: Academic Press. Na Internet prezentaciji predmeta Statistika u psihologiji (URL adresa: **http//:www.f.bg.ac.yu/**, slediti linkove **Odeljenja**, **Psihologija**, **Katedre**, **Statistika u psihologiji**) mogu se pronaći linkovi za prezentacije iz ove oblasti.

[23] O standardnoj grešci statistika će biti reči u glavi **VI**.

[24] O ocenjivanju parametara će biti reči u glavi **VI**.

- AS je veoma osetljiva na rezultate koji pokazuju ekstremna odstupanja u jednom smeru;[25]
- AS binarne varijable čije su kategorije označene brojevima 0 i 1 je jednaka proporciji jedinica u nizu.

Medijana (Mdn)

Medijana je mesto na brojnoj liniji ili u raspodeli iznad i ispod kojega je po 50% mera iz niza.

Uslovi: Bar ordinalna skala.

Računanje:

-Iz sirovih mera uredjenih po veličini odredjuje se mesto na kojem je Medijana po obrascu
$(n + 1) / 2.$

-Iz distribucije, tj. za kontinuirane nizove mera po obrascu:

$$Mdn = D + [(n/2 - F) / f]^* i$$

gde je D donja egzaktna granica razreda u kojem je Mdn; F- kumulativna frekvenca do razreda u kojem je Mdn; f - obična frekvenca u razredu u kojem je Mdn, a i predstavlja veličinu razrednog intervala. Razred u kojem je Mdn je onaj za koji je $cf > n / 2$ ili $cf = n / 2$.

Osobine:

- Matematičke operacije sa Mdn se izvode teže no sa aritmetičkom sredinom jer nije definisana standardnom jednačinom;
- Nestabilnija od AS od jednog do drugog uzorka;
- Zbir apsolutnih odstupanja svake mere od Medijane manji je nego od bilo koje druge vrednosti (ako se radi o diskretnim varijablama):

$$\sum_{i=1}^{n} |x_i - Mdn| = \min$$

- Nije osetljiva na ekstremne rezultate kao aritmetička sredina.

[25] U slučajevima kada postoje ekstremni rezultati, tj. autlajeri (engl. outlier) mogu se kao deskriptivne statističke mere koristiti tzv. **postriženi** (engl. trimmed) statistici, tj. statistici koji su izračunati iz distribucija iz kojih je 10% do 20% ekstremnih rezultata uklonjeno.

Primer:

I niz 67,59,75 M = 67.00; Mdn = 67.0.

II niz 1,59,75 M = 45.00; Mdn = 59.0.

Dakle, dva niza se razlikuju samo u jednom rezultatu koji je u drugom nizu veoma nizak. Medijane dva niza su dosta bliske dok im se aritmetičke sredine veoma razlikuju.

Mod

To je mera koja se najčešće javlja u nizu mera. Može se računati za bilo koje podatke. Može postojati više od jednog moda za isti niz podataka.
Računanje:

- Iz sirovih podataka: nalazi se mera koja je najčešća.

-Iz distribucije se može proceniti obrascem Mod = 3Mdn - 2M ili se može uzeti srednje mesto razreda sa najvećom učestalošću kao procena Moda.

Geometrijska sredina (G)

Geom. sredina je n-ti koren iz n umnožaka.
Uslovi za njeno korišćenje:
-Racio skala ;
-Sve mere moraju biti veće od 0.

Računanje:

$$G = \sqrt[n]{x_1 * x_2 * \ldots * x_n}$$

U oblasti psihofizike provere Fehnerovog zakona koriste geometrijsku sredinu. Ako je intenzitet draži u logaritamskom odnosu sa intenzitetom oseta onda bi stimulus koji je označen kao senzorno na sredini S_1 i S_2 trebalo da bude jednak njihovoj geometrijskoj sredini.

Zapamtite: Ako je distribucija **simetrična** onda su **AS, Mdn** i **Mod jednaki**.

IV.5.Mere varijabilnosti

Za statistički opis niza numeričkih podataka dobijenih merenjem odredjene pojave nije dovoljno navesti samo meru centralne tendencije datog niza. Potrebno je navesti i odgovarajuću meru varijabilnosti. Mere varijabilnosti su statističke mere koje izražavaju variranje ili raspršenje u nizovima numeričkih podataka. Kada bi sve mere u nizu bile jednake onda bi mera centralne tendencije bila tačan predstavnik svih ovih mera. Kako to najčešće nije slučaj, potrebno je znati koliko neka mera centralne tendencije dobro predstavlja niz iz kojega je izračunata.To se postiže tako što se uz svaku meru centralne tendencije po pravilu navodi i odgovarajuća mera varijabilnosti.

Primer: Data su dva niza mera :

I niz: 5 5 4 5 5 6 II niz: 1 4 8 10 5 2

Aritmetičke sredine oba niza su jednake (M niza I =M niza II =5.00), ali aritmetička sredina prvog niza mnogo bolje predstavlja podatke jer je raspon prvog niza 2, a drugog čak 9.(Standardna devijacija prvog niza je 0.63 a drugog čak 3.46).

Izbor mere varijabilnosti zavisi od :

-Vrste merne skale sa koje su podaci i

-Oblika distribucije

Raspon (R) je najgrublja mera varijabilnosti

Računa se po obrascu:

$$R = x_{max} - x_{min}$$

Osobine raspona:

- Zavisi samo od 2 rezultata u nizu;
- Osetljiv je na ekstremne rezultate.

Zanimljiv je za neke pojave:raspon cena na tržištu, raspon ličnih dohodaka...

Kvartilna devijacija (Q)

Kvartilna devijacija je, u stvari, polovina raspona koji se dobije eliminisanjem gornje i donje ekstremne četvrtine rezultata, tj. polovina interkvartilnog opsega.
Uslovi: Bar ordinalna merna skala.
Računa se najčešće za distribucije koje znatno odstupaju od normalne: asimetrične, polimodalne, diskontinuirane.
Obrazac za ovu meru je :

$$Q = (P_{75} - P_{25}) / 2$$

gde su P_{75} i P_{25} percentili (o percentilima se može pogledati u odeljku **IV.6.**).

Osobine:
Nije osetljiva na ekstremne mere u nizu.

Standardna devijacija (SD ili S)

Uslovi:
-bar intervalna merna skala i
-pretpostavka o normalnoj raspodeli mera u populaciji.

Stand. devijacija je kvadratni koren iz prosečnog kvadrata odstupanja svakog rezultata od AS svih rezultata.

Računanje:

$$S = \sqrt{\sum_{i=1}^{n}(x_i - M_x)^2 * n^{-1}}$$

Osobine SD:

- **Dodavanje** konstante svakoj meri u nizu **ne menja** SD niza;
- **Množenje** konstantom svake mere u nizu **povećava** SD **apsolutnu vrednost konstante puta;**
- Osetljiva je na ekstremne rezultate.

Varijansa (S^2)

Varijansa je aritmetička sredina (prosek) kvadrata odstupanja niza mera od njihove AS. Teorijski, varijansa je drugi centralni momenat (videti o momentimau odeljku **IV.7.**):
$\sigma^2 = E\,[X-E(X)]^2$
Računanje[26]:

$$S^2 = \sum_{i=1}^{n}(x_i - M_x)^2 * n^{-1}$$

ili

$$S^2 = \sum_{i=1}^{n}(x_i - M_x)^2 * (n-1)^{-1}$$

Moguće je za računanje varijanse koristiti i sledeći obrazac:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - M^2$$

Varijansa u matričnom obliku:

$$S^2 = (\mathbf{x}^t\mathbf{x} - \mathbf{x}^t\mathbf{P}\,\mathbf{x}) * n^{-1}$$

Prvi izraz $\mathbf{x}^t\,\mathbf{x}$ jednak je izrazu ispred minusa u brojiocu prethodnog obrasca, a $\mathbf{x}^t\,\mathbf{P}\,\mathbf{x}$ je jednak izrazu iza minusa u brojiocu. $\mathbf{P} = \mathbf{e}(\mathbf{e}^t\mathbf{e})^{-1}\mathbf{e}^t$ je projektor, matrica reda n x n čiji su svi elementi u ovom slučaju jednaki 1/n.
Prema tome, **varijansa je kvadrat standardne devijacije.**

Osobine varijanse:

- **Dodavanje** konstante svakoj meri u nizu **ne menja** varijansu niza;

[26] Prvi obrazac koristimo kada želimo samo da prikažemo varijabilnost uzorka a drugi kada na osnovu varijanse uzorka želimo da dobijemo nepristrasnu ocenu parametra (o oceni parametra na osnovu statistika se može pročitati u glavi **VI**).

- **Množenje** konstantom svake mere u nizu **povećava** varijansu niza **za kvadrat konstante** puta;
- Ima svojstvo aditivnosti (razbijanje u sastavne komponente) na kome počiva čitav niz postupaka nazvanih Analiza varijanse.

Varijansa je mera raspršenja rezultata oko AS i istovremeno greška koju pravimo kada na osnovu minimalnog broja informacija treba predvideti rezultat ispitanika. Ako su u vektoru **g** greške koje pravimo kada rezultate ispitanika koji su u vektoru **x** zamenimo aritmetičkom sredinom onda kao prosek kvadriranih grešaka dobijamo varijansu:

$\mathbf{g}^t\mathbf{g}\, n^{-1} = (\mathbf{x} - \mathbf{e}\ \mu)^t (\mathbf{x} - \mathbf{e}\ \mu)\, n^{-1} = (\mathbf{x}^t - \mu\, \mathbf{e}^t)(\mathbf{x} - \mathbf{e}\,\mu)\, n^{-1} = (\mathbf{x}^t\mathbf{x} - \mu\,\mathbf{x}^t\mathbf{e} - \mu\,\mathbf{e}^t\mathbf{x} + \mu^2\,\mathbf{e}^t\mathbf{e})\, n^{-1} = (\Sigma x^2 - \Sigma x\,\mu - \mu\Sigma x + \mu^2 n)\, n^{-1} = [\Sigma x^2 - 2\,\mu\Sigma x + (\Sigma x/n)^2 n]\, n^{-1} = [\Sigma x^2 - 2(\Sigma x/n)\Sigma x + (\Sigma x/n)^2 n]\, n^{-1} = [\Sigma x^2 - 2(\Sigma x)^2 n^{-1} + (\Sigma x)^2 n * n^{-2}]\, n^{-1} = [\Sigma x^2 - 2(\Sigma x)^2 n^{-1} + (\Sigma x)^2 n^{-1}]\, n^{-1} = [\Sigma x^2 - (\Sigma x)^2 n^{-1}]\, n^{-1}$

Poslednji izraz je očigledno varijansa (videti obrazac ranije).

Isto to možemo dobiti i preko projektora: [27]

$$\mathbf{g}^t\mathbf{g} * n^{-1} = (\mathbf{x} - \mathbf{P}\mathbf{x})^t(\mathbf{x} - \mathbf{P}\mathbf{x}) * n^{-1} = (\mathbf{x}^t\mathbf{x} - \mathbf{x}^t\mathbf{P}\,\mathbf{x}) * n^{-1}$$

Varijansa je, dakle, mera gubitka informacije do koje bi došlo kad bi se svakom ispitaniku umesto njegovog stvarnog rezultata pridružio prosek svih rezultata grupe kao predviđena vrednost za tog ispitanika. Zamena rezultata aritmetičkom sredinom je zapravo projekcija tih rezultata u prostor definisan činjenicom da ljudi postoje i to sa najmanjom mogućom greškom.
Isto tako, varijansa je i prosečna razlika svakog čoveka od svakog drugog:
Ako je distanca i-tog i j-tog ispitanika $d_{ij} = x_i - x_j$
i ako napravimo matricu distanci sa 0 u glavnoj dijagonali i sa vandijagonalnim elementima d_{ij} tada je varijansa:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij} / (n^2 - n)$$

Koeficijent varijacije (KV)

To je, u stvari, procenat standardne devijacije od aritmetičke sredine.

Uslovi:
-racio merna skala i
-pretpostavka o normalnoj raspodeli mera u populaciji.

---

[27] **Projektor (P)** je simetrična matrica koja pomnožena sama sobom daje sebe (idempotentna): $\mathbf{P}^t\mathbf{P} = \mathbf{P}^2 = \mathbf{P}$

Računa se po obrascu:

$$KV = (S / M) * 100.$$

Služi za poredjenje varijabilnosti dveju ili više grupa u pogledu iste merne karakteristike i za poredjenje varijablinosti iste grupe u pogledu dveju ili više mernih karakteristika.

Moguća su sledeća poredjenja:

a) Koja je varijabilnost veća, a koja manja;
b) Odredjivanje proporcije većeg varijabiliteta u manjem;
c) Odredjivanje koliko puta je veći onaj veći varijabilitet od manjeg.

U psihologiji se ponekad koristi koeficijent varijacije i kada imamo podatke sa intervalne merne skale.Tada treba voditi računa da je u tom slučaju možda smisaono samo poredjenje pod a) i samo ako poredimo dve ili više grupa po jednoj mernoj karakteristici.

IV.6. Percentili i percentilni rangovi

Služe za:

1.Računanje nekih drugih statističkih mera, na primer kvartilne devijacije.
2.Određivanje položaja pojedinca u grupi kojoj pripada
3.Kvantitativno klasifikovanje, kategorizaciju i selekciju ispitanika.

Percentil (P_p) je mesto u raspodeli ispod kojega je odredjeni postotak slučajeva.
Decili su P_{10}, P_{20}, P_{30}....Kvartili su P_{25}, P_{50}, P_{75}.
Računanje:

$$P_p = D + [(Pn - F) / f]* i$$

gde oznake imaju isto značenje kao kod Mdn a Pn=(P n)/100. Razred u kojem je percentil je onaj u kojem je kumulativna frekvenca cf > Pn ili cf = PN

Percentilni rang ($PR_{(x)}$) je mesto koje zauzima ispitanik na procentnoj skali s obzirom na meru koju ima, a u odnosu na grupu kojoj pripada.
Percentilni rang se računa na sledeći način:

$$PR_{(x)} =(100/n) (F+[(x- D)/i]* f)$$

Oznake u prethodnom obrascu imaju isto značenje kao kod percentila, samo je u pitanju razred u kojem je mera x, a ne razred u kojem je percentil.

IV.7.Momenti i momentni koeficijenti

Moment [28]za potenciju p i tačku a (ili moment reda p oko tačke a) definisan je kao:[29]

$$\mu_a^p = \sum_{i=1}^{n}(x_i - a)^p$$

Momentni koeficijent je:

$$\varphi_a^p = \sum_{i=1}^{n}(x_i - a)^p * N^{-1}$$

Prvi moment oko nule je:

$$\mu_{a=0}^{p=1} = \sum_{i=1}^{n} x_i$$

Aritmetička sredina je momenti koeficijent reda 1 definisan oko 0:

$$\varphi_{a=0}^{p=1} = \sum_{i=1}^{n} x_i * N^{-1}$$

Centralni momenti

Centralni momenti su momenti reda p definisani oko aritmetičke sredine.

Suma kvadrata odstupanja rezultata od aritmetičke sredine je drugi centralni moment:

$$\mu_{a=\varphi_0^1}^{p=2} = \sum_{i=1}^{n}(x_i - \mu)^2$$

Varijansa je drugi centralni momentni koeficijent:

$$\varphi_{a=\mu}^{p=2} = \sum_{i=1}^{n}(x_i - \mu)^2 * N^{-1} = \sigma^2$$

Treći centralni momentni koeficijent

$$\varphi_{a=\mu}^{p=3} = \sum_{i=1}^{n}(x_i - \mu)^3 * N^{-1} = \mu_3$$

Treći centralni momenti koeficijent se koristi za računanje koeficijenta asimetrije (skewness):

$$\gamma = \left(\sum_{i=1}^{n}(x_i - \mu)^3 .* N^{-1}\right) / \sigma * \sigma^2$$

---

[28] Obrasci za momente i momentne koeficijente koji su ovde dati odnose se na populaciju pa su, stoga, oznake parametara napisane grčkim slovima, a veličina populacije je označena slovom N.

[29] Termin moment potiče iz mehanike. Ako se sila f_1 primeni na rastojanju x_1 od početne tačke onda je f_1x_1 momenat sile. Ako se druga sila f_2 primeni na rastojanju x_2 ukupni momenat je $f_1x_1 + f_2x_2$. Ako kvadriramo distance dobijamo drugi moment itd.

Skjunis za normalnu ditribuciju jednak je nuli, negativan je za negativno asimetričnu, a pozitivan za pozitivno asimetričnu distribuciju. U intervalu od -0.5 do 0.5 je umerena asimetrija, a izvan toga je znatna. Distribucija koeficijenta zakrivljenosti je normalna.

Četvrti centralni momentni koeficijent

$$\varphi_{a=\mu}^{p=4} = \sum_{i=1}^{n}(x_i - \mu)^4 N^{-1} = \mu_4$$

Četrvrti momentni koeficijent se koristi za računanje koeficijenta izduženosti (Kurtosis):

$$\lambda = ((\sum_{i=1}^{n}(x_i - \mu)^4 * N^{-1}) / \sigma^2 * \sigma^2) - 3$$

Kurtozis za normalnu ditribuciju jednak je nuli. Ako je kurtozis veći od 0 distribucija je izdužena (leptokurtična), a ako je manji od 0 raspodela je spljoštena (platikurtična). Ako je kurtozis manji od -1.2 distribucija je U tipa.[30]

****

Izračunavanje pojedinih mera centralne tendencije, mera varijabilnosti, percentila, skjunisa i kurtozisa u SPSS programu moguće je uraditi pomoću menija **Summarize** i podmenija **Frequencies** (u poddijalogu **Frequencies: Statistics**) i **Descriptives** (u poddijalogu **Descriptives: Options**) odabirom odredjene opcije.
Ako ove mere računamo u proceduri **Frequencies** a ne želimo da, pri tome, dobijemo u ispisu i raspodelu učestalosti potrebno je u prozoru za dijalog isključiti (automatski inače uključenu) opciju **Display frequency tables** koja se nalazi ispod levog spiska varijabli.

Primer: Ispis mera centralne tendencije, mera varijablilnosti i mera oblika distribucije iz procedure **Frequencies**:

```
EXTRA      AJZENK_EKSTRAVERZIJA

Mean          13.642      Median        14.000      Std dev        4.118
Kurtosis       -.306      S E Kurt        .280      Skewness       -.599
S E Skew        .140

Percentile     Value      Percentile     Value      Percentile     Value

  25.00       11.000        50.00       14.000        75.00       17.000

Valid cases      302      Missing cases      14
```

Ispis iz procedure **Descriptives**:

```
Number of valid observations (listwise) =       302.00

Variable Mean  Std Dev Kurtosis S.E. Kurt Skewness S.E. Skew Valid N

EXTRA   13.64   4.12     -.31       .28      -.60       .14       302
```

[30] Skjunis i kurtozis se u programu SPSS računaju na opisane načine. Pri tumačenju rezultata iz ovog programa ne treba uzimati u obzir preporuke iz Dragićevićeve knjige Statistika za psihologe jer su u njoj date vrednosti Skjunisa i Kurtozisa računatih preko drugačijih obrazaca.

Uočimo da u proceduri Descriptives ne možemo dobiti percentile (pa ni Medijanu). S.E. Kurt i S.E. Skewness su standardne greške Skjunisa i Kurtozisa koje program automatski računa ako zatražimo Skjunis i Kurtozis. Ove standardne greške mogu poslužiti u testiranju normalnosti raspodele varijable ekstroverzije u populaciji (o tome će biti reči kasnije).

****

Računanje percentilnih rangova u SPSS programu može se izvesti tako što se rezultati na varijabli ranguju (procedura **Transform/Rank Cases**), a zatim se na osnovu varijable sa rangovima (čije ime se dobija tako što program dodaje slovo **r** ispred imena izvorne varijable) računaju percentilni rangovi komandom **Transform/Compute**. To se radi tako što se od ranga koji odgovara određenoj meri oduzme 0.5, i zatim ova razlika podeli ukupnim brojem rezultata, tj. ispitanika čije rezultate imamo. Dobijenu vrednost moguće je pomnožiti sa 100 da bi se percentilni rang mere iskazao procentom. Na primer, ako su u varijabli REXTRA iz prethodnog primera rangovi za varijablu ekstroverzija, komandom

COMPUTE PR_EXTR = [100*(REXTRA - 0.5)]/302

dobili bismo percentilne rangove za sve rezultate (ispitanike) koji bi bili upisani u varijabli PREKSTRA na kraju aktivnog fajla sa podacima.

****

IV.8. Neke važne transformacije podataka na kvantitativnoj varijabli
Centriranje i standardizacija varijable

Originalne podatke sa kvantitativne varijable X možemo transformisati na praktično neograničen broj načina. Najčešće korišćene transformacije su:

1. **Linearne** - podaci se transformišu u drugi oblik linearnom funkcijom $f(x) = a + bx$.
2. **Nelinearne** - podaci se transformišu u drugi oblik nelinearnim funkcijama. Na primer: $f(x) = e^{bx}$.
3. **Monotone** - podaci se transformišu tako da se očuva originalni redosled. Transformacije $f(x)$ pri čemu važi da ako je $x_i \leq x_j$ onda je $f(x_i) \leq f(x_j)$ su takvog tipa.

****

Za različite tipove transformacija podataka u SPSS programu može poslužiti meni **Transform** i podmeni **Compute** (videti objašnjenje komande **Compute** u odeljku **IV.3.**)

****

Centriranje rezultata

Centriranje rezultata na varijabli sastoji se u tome da se svaki rezultat iskaže kao odstupanje od aritmetičke sredine svih rezultata:

$$d_i = (x_i - M)$$

Devijaciona mera ili **centrirani skor** (d_i) je, dakle, rezultat iskazan kao odstupanje izvornog, netransformisanog ili sirovog skora od aritmetičke sredine niza kojem skor pripada.

Zapamtite: **Centrirani rezultati imaju aritmetičku sredinu jednaku 0 a varijansu jednaku varijansi izvornih rezultata.**

Standardizacija

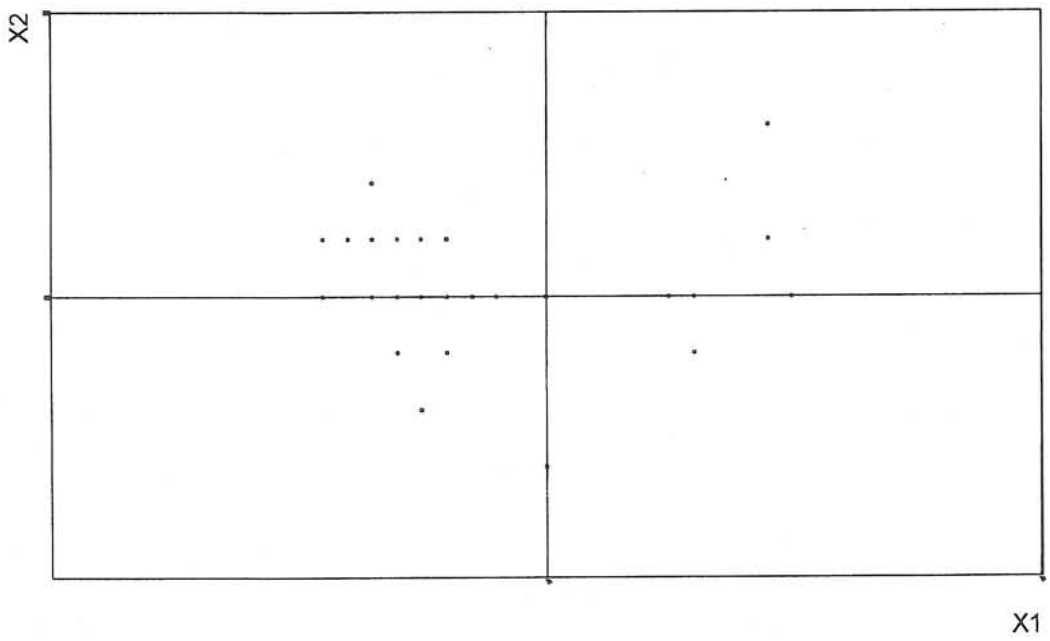
Standardizacija je transformacija rezultata u oblik u kojem podaci ne zavise od mernih jedinica:[31]

$$z_i = (x_i - M) / S$$

Za standardizovane mere važe sledeće osobine:

1. $\sum_{i=1}^{n} z_i = 0$ jer je $\sum_{i=1}^{n} (x_i - M) = 0$
2. $\sum_{i=1}^{n} z_i^2 / n = 1 \Leftrightarrow \sum_{i=1}^{n} z_i^2 = \mathbf{z}^t\mathbf{z} = n$

Grafik 1: Grafički prikaz različitog varijabiliteta dveju varijabli



[31] Standardizovanjem podataka mi zapravo izravnavamo svaku varijablu u skladu sa njenim varijabilitetom. Kao što se na Grafiku 1 može videti raspršenje tačaka na x_1 koordinati je šire no na koordinati x_2, tj. više je inherentne varijabilnosti u x_1 nego u x_2. (Ove dve varijable kako se iz crteža vidi nisu povezane, tj. nisu u korelaciji. O korelaciji će biti reči u glavi **XI**). Budući da je širenje tačaka pretežno duž x_1 dimenzije postaje razumljivo da se za istu vrednost, pri računanju distance od neke tačke, x_2 koordinata više ponderiše. Velike vrednosti na x_2 su neverovatne. To upravo radimo pri standardizaciji varijabli: vršimo ponderisanje sa $1/\sigma_1$i $1/\sigma_2$ čime dobijamo nove koordinate z_1 i z_2. Distanca tačaka x_1 i x_2 od izvora (koordinatnog početka) može se izraziti preko njihovih standardizovanih koordinata $z_1=x_1/\sigma_1$ i $z_2=x_2/\sigma_2$:

$$|| X' || = || X' - 0 || = [(z_1 - 0)^2 + (z_2 - 0)^2]^{1/2} = [(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2]^{1/2}$$

Zapamtite:

- **Standardizovani** rezultati imaju **aritmetičku sredinu** jednaku **0,** a **varijansu** jednaku **1.**
- Rezultati u **standardnom normalnom** obliku imaju **aritmetičku sredinu** jednaku **0,** a **varijansu** jednaku **1/n.**

Transformacija rezultata u standardni normalni oblik:

Standardni normalni oblik rezultata dobija se deljenjem standardizovanih rezultata kvadratnim korenom veličine uzorka, tj. broja rezultata:

$$z_n = z\, n^{-1/2}$$

Zbir kvadriranih rezultata koji su transformisani u standardni normalni oblik, tj. dužina vektora varijable u kojoj su ovakvi rezultati jednaka je jedinici[32]:

$$z_n^{t} z_n = 1.$$

****

Centriranje, standardizovanje varijable i transformacija u standardni normalni oblik u SPSS programu se mogu izvesti naredbom **Compute** (meni **Transform**), a standardizacija se osim na taj način može automatski napraviti i uključivanjem opcije **Save standardized values as variables** (klikom na kvadratić ispred te opcije) u proceduri **Descriptives** menija **Statistics**. Na kraju ispisa iz procedure **Descriptives** (ako je ukljucena opcija za standardizovanje varijabli) program obaveštava korisnika o imenima standardizovanih varijabli koje su upisane na kraju aktivnog fajla sa podacima.

****

IV.9. Grafičko prikazivanje jedne kvantitativne varijable

Jedna stara kineska poslovica kaže: "Slika vredi koliko i hiljadu reči".Grafički predstavljeni podaci postaju pregledniji i bolje se razumeju. Jedan pogled na grafik ponekad više govori od dugotrajnog pregledanja kolona sa cifarskim podacima.
Grafički prikaz raspodele omogućuje i poređenje dveju raspodela.

Četiri osnovne "klasične" metode su:

1. Poligon učestalosti. Da bi se konstruisao ovaj grafik, na apscisi se nanose mere za varijablu u intervalima razreda, a na ordinatu frekvence. Poligon dobijamo tako što pravim linijama povežemo tačke čije su koordinate srednje mesto razreda i učestalost u datom razredu.

2. Histogram - poligon stubaca. Apscisa i ordinata su iste kao kod poligona učestalosti a histogram nastaje tako što iznad segmenta [D, G] na apscisi, koji je definisan donjom (D) i gornjom (G) egzaktnom granicom razreda, konstruišemo pravougaonik čija visina

[32] Dužina vektora je kvadratni koren zbira kvardriranih elemenata tog vektora. Budući da je u ovom slučaju zbir kvadriranih elemenata vektora jednak jedinici, to je i njegova dužina jednaka jedinici.

odgovara frekvenci za dati razred. Na taj način, učestalost je prikazana površinom stupca, a frekvence su prikazane kao da su podjednako raspodeljene u datom intervalu.

3. Grafik kumulativnih frekvencija. Ovaj grafik dobijamo tako što pravim linijama povežemo tačke čije su koordinate gornja egzaktna granica razreda i kumulativna frekvenca za dati razred.

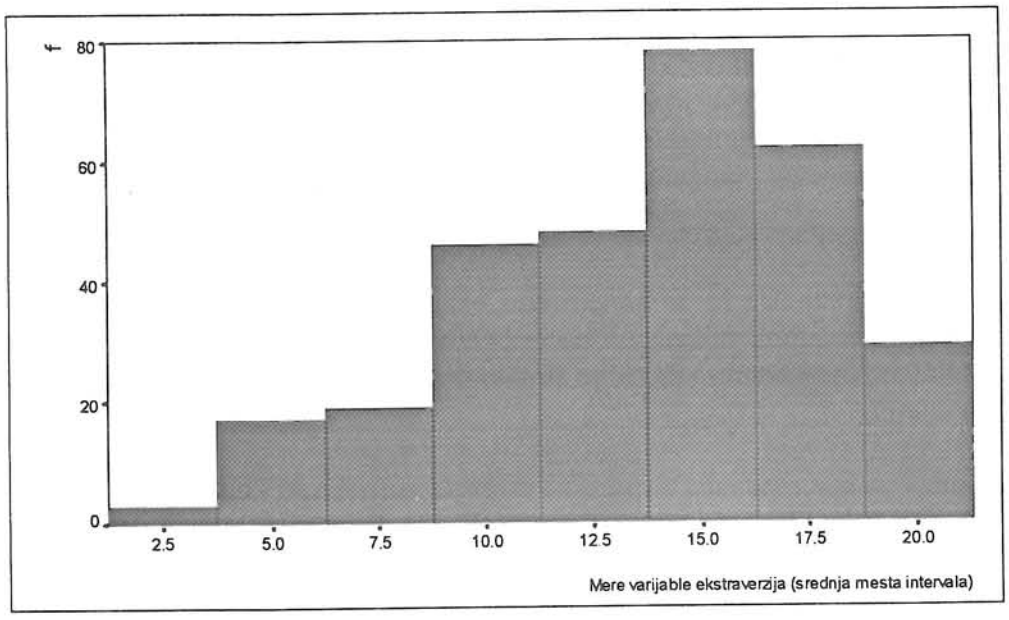
4. Kriva kumulativnih procenata - Ogiva. Ovaj grafik dobijamo tako što pravim linijama povežemo tačke čije su koordinate gornja egzaktna granica razreda i relativna kumulativna frekvenca za dati razred.[33]

U ovom tekstu će biti objašnjen samo postupak konstruisanja Histograma u SPSS-u.

****

U SPSS-u je najbolje (premda postoje i druge mogućnosti) crtati histogram odabirom **Graphs/Histogram**. U prozoru za dijalog **Histogram** ime varijable koja se želi pikazati histogramom treba prebaciti iz spiska u levom okviru u polje **Variable**. Klikom na dugme **Titles...** moguće je u poddijalogu **Titles** dati naslov crteža. Klikom na dugme **Continue** ovog poddijaloga vraćamo se (po unetom naslovu) u dijaloški prozor **Histogram**. Klikom na dugme **OK** završavamo definisanje histograma. Moguće je, uključivanjem opcije **Display normal curve** u dijaloškom prozoru **Histogram** zatražiti da se normalna kriva superponira preko histograma. Grafici koji nastaju u starijim verzijama SPSS programa prikazuju se u prozoru **Chart Carousel**. Ovde je moguće odustati od grafika i napustiti prozor **Chart Carousel** klikom na dugme **Discard** ili (klikom na dugme **Edit**) ući u editor grafika. Editor grafika ima svoj meni (**File**, **Edit**, **Gallery...**) pomoću kojega se može grafik doterivati i dalje menjati.
U novijim verzijama SPSS-a u grafički editor se ulazi **dvostrukim** klikom na grafik koji se nalazi u okviru prozora za ispis.

Primer : Histogram za varijablu EKSTRA čiju smo raspodelu učestalosti pokazali ranije izgledao bi ovako:



[33] Detaljnija objašnjenja postupka konstruisanja ovih grafika data su u knjizi Č. Dragićevića Statistika za psihologe.

Iz grafika treba uočiti dve bitne stvari:

1. Apscisa (rezultati ispitanika na varijabli su u intervalima razreda) i ordinata (učestalost mera po razredima- **f**) su označene odgovarajućim nazivima.To je obavezno u svakom grafičkom prikazu podataka!
2. Prikazani su samo rezultati ispitanika za koje imamo podatke, tj. n = 302.

****

Za eksploratornu analizu podataka veoma su korisni i sledeći grafički prikazi:

1. Prikaz "Stabljika sa lišćem" (stem and leaf plot) i

2. Crtež "kutija sa kanapima" (Box and Whisker plot)

1. Crtež **stabljika sa lišćem** nije pravi grafički prikaz. On se sastoji iz dva skupa brojeva razdvojenih vertikalnom linijom. Svaki rezultat se deli na stabljiku (vodeću cifru) i list (prateću cifru). Sa leve strane su "vodeće cifre", a sa desne strane svake vodeće cifre su preostale ("prateće") cifre određenih mera (čija je to vodeća cifra). Tako npr., ako su rezutati na nekoj varijabli 11, 13, 23, 23, 23, 24, 26, 29, 31, 33 prikaz će izgledati ovako:

Steam	Leaf
1	13
2	333469
3	13

****

Da bi se napravio prikaz **Stabljika sa lišćem** u SPSS programu treba odabrati **Statistics/Summerize/Explore** i u prozoru za dijalog **Explore** prvo odabrati (kliknuti na dugme ispred opcije tako da se zatamni) opciju **Plots** u okviru **Display**. U okvir **Dependent list** prebaciti iz levog spiska ime kvantitativne varijable za koju želimo ovakav prikaz. Potom pređemo u prozor poddijaloga **Explore: Plots** (klikom na dugme **Plots**) prozora **Explore** i u okviru **Descriptive** uključimo (mada je obično već uključena po default-u programa) opciju **Stem-and-leaf**. U okviru **Boxplots** (ako se ne želi istovremeno i boxplot) treba odabrati opciju **None**. Ukoliko postoji veliki broj slučajeva ili veliki raspon rezultata onda je ispod prikaza data širina intervala (**Stem width:** *broj*) između sukcesivnih vodećih brojeva i infomacija o tome koliko je slučajeva približno prikazano jednom pratećom cifrom (**Each leaf** *broj* **case(s)**). Pored toga u koloni **Frequencies** sa leve strane prikaza date su učestalosti pojedinih grupa vrednosti sa istom vodećom cifrom.

Primer:

"Stabljika sa lišćem" za varijablu ekstroverzija na uzorku beogradskih studenata (n=302):

```
EXTRA     AJZENK_EKSTRAVERZIJA
 Valid cases: 302   Missing cases:  14   Percent missing:     4.4

 Frequency     Stem &  Leaf
     2.00 Extremes     (2.0)
     1.00          3 .  0
     4.00          4 .  0000
     7.00          5 .  0000000
     6.00          6 .  000000
    11.00          7 .  00000000000
     8.00          8 .  00000000
    12.00          9 .  000000000000
    14.00         10 .  00000000000000
    20.00         11 .  00000000000000000000
    23.00         12 .  00000000000000000000000
    25.00         13 .  0000000000000000000000000
    23.00         14 .  00000000000000000000000
    30.00         15 .  000000000000000000000000000000
    25.00         16 .  0000000000000000000000000
    35.00         17 .  00000000000000000000000000000000000
    27.00         18 .  000000000000000000000000000
    17.00         19 .  00000000000000000
    12.00         20 .  000000000000
 Stem width:      1
 Each leaf:       1 case(s)
```

Uočimo da je u ovom prikazu sadržano ono što imamo u raspodeli učestalosti (kolona **Frequency**) kao i ono što vidimo iz histograma (listovi, tj. 0 grade oblik raspodele). Na dnu prikaza objašnjeno je kolika je širina korena (kolika je razlika sukcesivnih vrednosti u koloni **Stem**) i koliko rezultata je prikazano jednom cifrom u desnom odeljku **leaf**. U ovom slucaju rezultati u koloni **Stem** razlikuju se za jednu jedinicu (**Stem width: 1**) a svaka 0 u delu za lišće stoji za jednog ispitanika (**Each leaf: 1 case(s)**).

****

2. Kutija ili sanduče na kanapima (ili Boxplot)

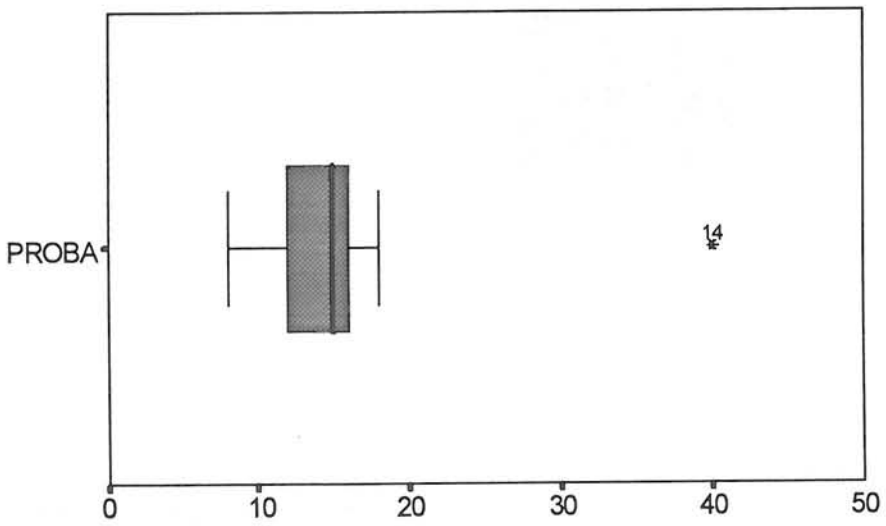
Crtež "kutija ili sanduče na kanapima" daje grafički prikaz distribucije preko tri komponente:

1. Središnja linija u sandučetu prikazuje centralnu tendencu, najčešće Medijanu. Na osnovu položaja ove linije unutar sandučeta može se zaključiti da li je distribucija simetrična (linija je na sredini sandučeta) ili asimetrična (linija je bliže levoj ili desnoj ivici sandučeta).
2. Sanduče, čije ivice približno odgovaraju trećem i prvom kvartilu (utoliko više što je uzorak veći). Lokacija rezultata na kojima će biti gornja i donja ivica sandučeta se računaju tako da se na lokaciju medijane [(n+1)/2] doda jedinica i dobijeni rezultat podeli sa 2. Ivice se zatim ucrtavaju tako da odgovaraju k-tom najmanjem i k-tom najvećem rezultatu gde je k lokacija (zaokružena na ceo broj) ivica dobijena na opisani način. H-opseg je raspon između gornje i donje ivice sandučeta.
3. Kanapi , tj. linije sa jedne i druge strane sandučeta ucrtavaju se do najniže i najviše okolinske vrednosti. Donja okolinska vrednost je najmanji rezultat koji je veći ili

jednak donjoj ogradi, a gornja okolinska vrednost je najveći rezultat koji je manji ili jednak gornjoj ogradi. Donja ograda se računa tako što se od donje ivice oduzme vrednost koja je jednaka 1.5*H-opsega, a gornja ograda se dobija tako što se na gornju ivicu doda 1.5*H-opseg.

4. Zvezdicama se ucrtavaju rezultati koji su veći od gornje i manji od donje okolinske vrednosti. Tako se vizuelno uočavaju ekstremne vrednosti –autlajeri (outliers)

Primer Za sledeći niz brojeva (rezultati na varijabli PROBA): 8,10,12,12, 14,14,15,15,15,15, 16,17,18,40 ovaj grafik bi izgledao ovako:

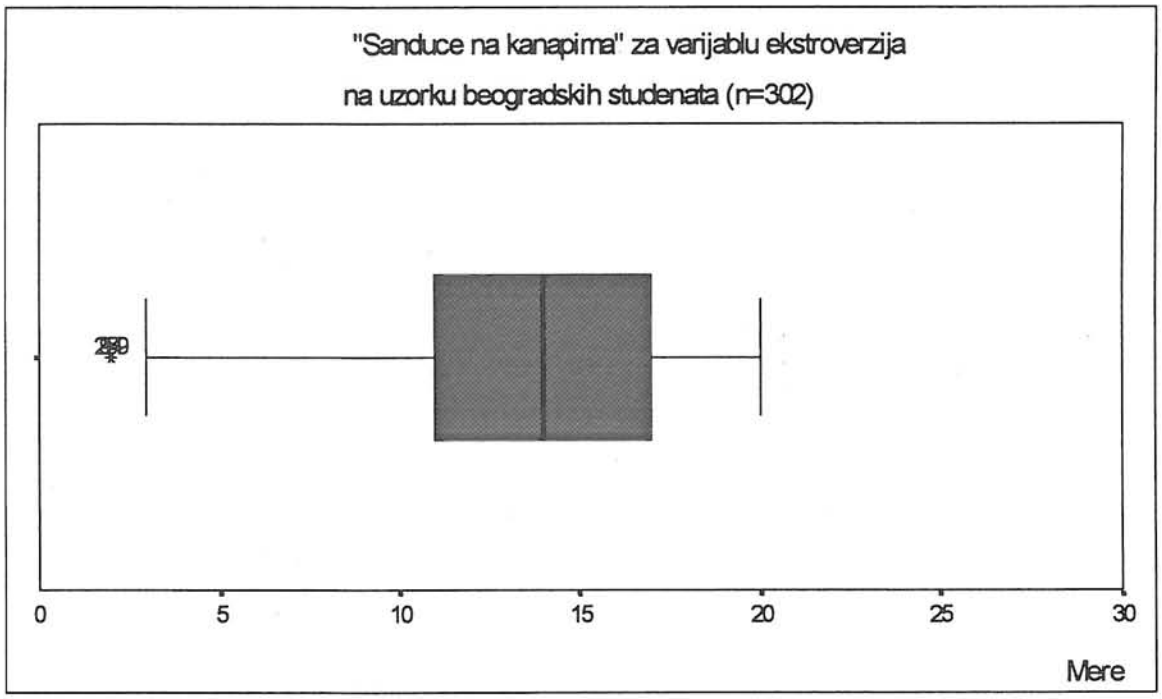


Lokacija medijane je (14+1)/2 pa je medijana na sredini između sedmog i osmog rezultata i iznosi 15. Medijana je na grafiku prikazana kao linija koja preseca kutiju. Iz činjenice da ova linija u prikazanom primeru nije podjednako udaljena od obe ivice kutije možemo zaključiti da raspodela nije simetrična. Lokacija ivica kutije je (7.5+1)/2 pa je gornja ivica kutije četvrti rezultat počev od najvišeg (to je ovde skor 16) a donja ivica četvrti rezultat počev od najnižeg (u ovom slučaju 12). Gornja i donja ivica su veoma blizu prvog (ovde je on 12) i trećeg kvartila (koji je ovde 16.25) pa se može reći da kutija predstavlja interkvartilni raspon. H-opseg je 16-12 = 4 pa je donja ograda 12 - 1.5*4 = 6 a gornja 16 + 1.5*4=22. Donja okolinska vrednost je 8 a gornja 18. Rezultat 40 je prikazan zvezdicom i označen svojim rednim brojem u nizu prikazanih rezultata. Očigledno se radi o outlier-u, tj. ekstremnom rezultatu.

****

Da bi se jedna kvantitativna varijabla prikazala crtežom "kutija na kanapima" u SPSS programu je najbolje odabrati **Graphs/Boxplot**, a u prozoru za dijalog **Boxplot** kliknuti na ikonu **Simple,** u okviru **Data in charts are** uključiti opciju **Summaries of separate variables** i potom klikom na dugme **Define** preći u prozor poddijaloga **Define Simple Boxplot: Summaries of Separate Variables**. U okvir **Boxes Represent** ovog poddijaloga treba ubaciti ime varijable koja se prikazuje, a u polje **Label Cases by** moguće je definisati kako će entiteti sa ekstremnim rezultatima biti na grafiku označeni (pod uslovom da "varijabla" sa oznakama entiteta postoji u fajlu sa podacima).

Primer:



Iz grafika se lako uočava negativna asimetričnost distribucije (Kutija, koja predstavlja srednjih 50% slučajeva je pomerena udesno, a Medijana nije na sredini "kanapa"); a iz prethodnog ispisa već znamo da je Skjunis negativan. Isto tako, dva ispitanika sa izuzetno niskim rezultatima (autlajeri, engl. outliers) su prikazani zvezdicom levo od ivice "kanapa". S obzirom na to da su mere nanete na apscisi, moguće je približno odrediti da je Medijana blizu mere 14 (iz ispisa znamo da je jednaka 14).

****

V. STATISTIČKI OPIS I GRAFIČKI PRIKAZ JEDNE KATEGORIČKE (NOMINALNE) VARIJABLE

Struktura podataka

$E = \{ e_i : i = 1,....,n\} \subset P$ je slučajni uzorak izvučen iz populacije;
$Q = \{ q_k ; k = 1,2,..., g\}$ je kategorička varijabla koja se sastoji od g kategorija.

Cilj operacije "merenja" kategoričke varijable je definisanje odnosa ispitanika e_i prema kategorijama q_k varijable Q, tako što se svakom ispitaniku pridruži oznaka kategorije kojoj pripada. Simbolički tbismo mogli to označiti ovako:

$$e_{ik} = o_k \mid e_i \in q_k$$

gde je o_k oznaka za kategoriju **k** koja se pridružuje ispitaniku ako pripada k-toj kategoriji varijable Q. Opis skupa E na jednoj kategoričkoj varijabli, tj. na skupu Q može se predstaviti kao Dekartov proizvod ova dva skupa, a operacija kojom se to postiže zove se **klasifikovanje** ili **kategorisanje** entiteta. Klasifikovanje predstavlja po nekim autorima najniži nivo "merenja".

V.1.Definisanje i kodiranje nominalne (kategoričke) varijable

Nominalna varijabla je skup $Q = \{ q_k ; k = 1,2,..., g\}$ za $g \geq 2$, ako su ispunjeni ovi uslovi:

1. $q_k \cap q_j = \varnothing \mid k \neq j$; Presek različitih kategorija je prazan skup, tj. varijabla se sastoji od distinktnih (neprepokrivajućih) kategorija;

2. $\bigcup_1^g q_k = Q$; Unija svih kategorija jednaka je nominalnoj varijabli;

3. Varijabla mora biti tako konstruisana da je svaki entitet moguće svrstati u neku kategoriju, tj. dodeliti mu oznaku neke kategorije. Ipak, neke kategorije mogu biti prazne. Dakle, klasifikacija entiteta po kategorijama nominalne varijable mora biti isključiva i iscrpna.

Nominalna varijabla se može kodirati na dva načina:

1.Kao **faktor** - svakoj kategoriji se dodeli neka (najčešće cifarska) oznaka. Zatim se svakom ispitaniku pridruži oznaka kategorije nominalne varijable (faktora) kojoj pripada:

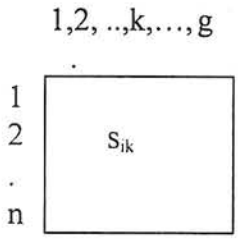
$$e_{ik} = o_k \mid e_i \in q_k$$

Na primer, na varijabli POL oznaka za muško može biti 1, a za žensko 2. U tom slučaju kao strukturu podataka za ovako kodiranu nominalnu varijablu dobijamo niz (vektor) reda nx1 u kojem su za svakog ispitanika unete oznake kategorijalne pripadnosti. U gornjem primeru vektor POL bi sadržavao cifre 1 i 2.

2.U **kompletni disjunktni oblik**. Kodiranje nominalne varijable u kompletni disjunktni oblik izvodi se tako što se konstruiše matrica **S**, takva da je:

$$\mathbf{S}=(s_{ik})=\begin{cases}1, & e_i \in q_k \\ 0, & e_i \notin q_k\end{cases} \quad i=1,\cdots,n, \quad k=1,\cdots,g$$

Dakle, na ovako kodiranoj kategoričkoj varijabli ispitanik e_i dobija oznaku 1 ako pripada datoj kategoriji, a oznaku 0 za sve druge kategorije, tj. za kategorije kojima ne pripada. Na taj način, kao strukturu podataka, dobijamo selektorsku ili indikatorsku matricu $\mathbf{S}=(s_{ik})$ reda **n** x **g**:



Jedna nominalna varijabla, na taj način, predstavlja **skup** od g ortogonalnih binarnih vektora.[34] Dok je na kvantitativnoj varijabli mera ispitanika definisana jednim brojem, dotle se kategorijalna pripadnost ispitanika na jednoj nominalnoj varijabli može definisati preko g binarnih vrednosti.

Jedna nominalna varijabla zapravo proizvodi višedimenzionalnu strukturu (broj dimenzija jednak je broju kategorija nominalne varijable). Kada se ovo ima u vidu postaje jasno zašto su postupci za analizu kategoričkih varijabli daleko složeniji od postupaka za analizu kvantitativnih varijabli.

V.2.Statistički opis kategoričke varijable

Prebrojavanje slučajeva po kategorijama kategoričke varijable:

oznake kategorija	frekvence
1	f_1
.	.
k	f_k
.	.
g	f_g

Ako je reč o jednostavnom slučajnom uzorku relativne frekvence (f_k/n) po kategorijama su ocene verovatnoća (p_k).

---

[34] Dva vektora x i y koji imaju po n elemenata su ortogonalna ako je njihov skalarni proizvod jednak nuli: $x'y=\sum_{i=1}^{n} x_i y_i = 0$.

Ako se ovo prebrojavanje iskazuje preko matrice **S**, onda je $\mathbf{F} = \mathbf{S}^t\mathbf{S}$, a dijagonalni elementi matrice **F** su frekvence pojedinih kategorija, tj. $(f_{kk}) = f_k$.
Operacijom $\mathbf{F}^*(1/n)$ dobijamo matricu $\mathbf{P} = (p_{kk})$, čiji su elementi relativne frekvence po kategorijama kategoričke varijable a, ukoliko je uzorak prost slučajni uzorak, relativne frekvence su ocene verovatnoća pojedinih kategorija.

Entropija nominalne varijable

Jednaka je informaciji koju nosi nominalna varijabla i analogna je varijansi kvantitativnih varijabli. Govori o tome koliko informacija nosi činjenica da znamo kategorijalnu pripadnost ispitanika.

Klasični načini računanja entropije:

1. $H = -\sum_{k=1}^{g} p_k \log_{10} p_k$ (u hartlijima);
2. $H = -\sum_{k=1}^{g} p_k \log_2 p_k$ (u bitima)

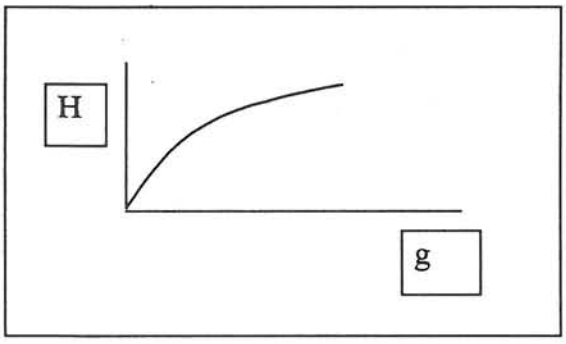
U statistici se entropija uobičajeno računa na sledeći način:

$H = -\sum_{k=1}^{g} p_k \log_e p_k$ (u prirodnim jedinicama)

Entropija kategoričke varijable zavisi od:

1. verovatnoće da se nešto dogodi i dostiže maksimum za $p_k = p$ $(\forall k)$;
2. broja kategorija koje varijabla ima (funkcija zavisnosti ima parabolični oblik - slika 6)

Slika 6 : Odnos između broja kategorija (g) na varijabli i entropije varijable (H) -šematski prikaz



Entropija se, radi dalje analize, može razbijati u delove.

****

Raspored učestalosti po kategorijama nominalne varijable se u SPSS programu pravi menijem **Statistics/Summarize/Frequencies** sa uključenom opcijom **Display frequency tables**. Budući da je reč o kategoričkoj varijabli nema smisla pri tom koristiti opciju **Statistics** (osim za određivanje Moda) u dijaloškom prozoru **Frequencies**. U ispisu se dobijaju oznake kategorija (kolona **Value**) i njihove učestalosti (**Frequency**), relativne učestalosti (sa nedostajućim podacima - kolona **Percent** i bez uzimanja u obzir nedostajućih podataka pri određivanju ukupnog n - kolona **Valid Percent**).
Transformisanje (rekodiranje) vrednosti nominalne varijable radi se komandom **Recode**. Ukoliko su oznake kategorija na nominalnoj varijabli unete kao alfanumerici (ili slovni nizovi) onda se komandom **Automatic Recode** mogu pretvoriti u cifarske oznake. Ove komande su objašnjene u odeljku **IV.3.**

Primer: Raspodela rukosti 413 ispitanika po kategorijama:

```
DOMINRUK  dominantna ruka

                                          Valid      Cum
Value Label                Value  Frequency Percent  Percent  Percent

leva                           1        124    30.0     30.0     30.0
obe                            2         75    18.2     18.2     48.2
desna                          3        214    51.8     51.8    100.0
                                      -------  -------  -------
                           Total          413    100.0    100.0

Valid cases      413      Missing cases      0
**********
```

V.3.Grafičko predstavljanje kategoričke varijable

Za grafičko predstavljanje nominalne varijable koriste se najčešće štapićasti i pitasti dijagram.

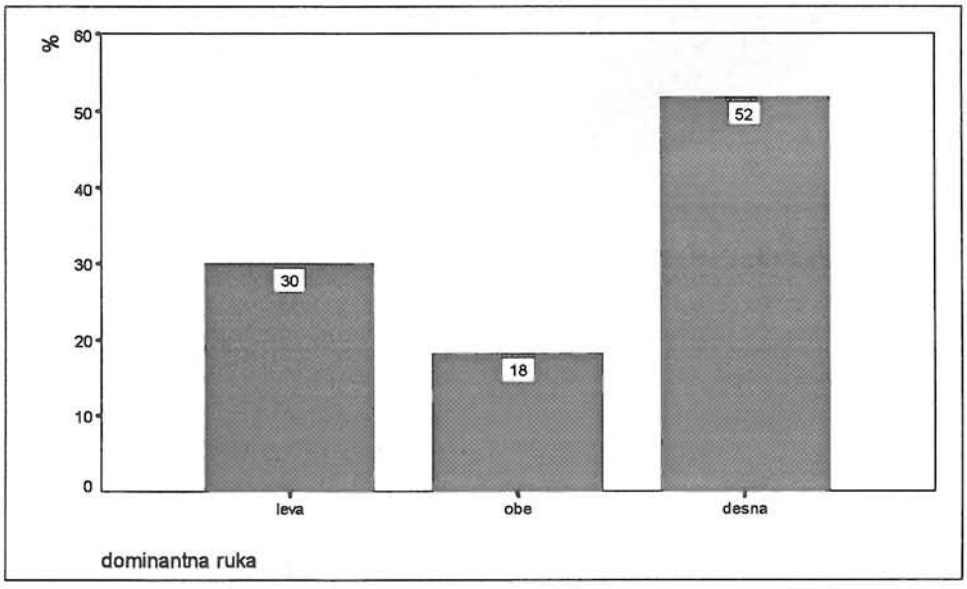
****

Štapićasti dijagram (Barchart)

Grafički prikaz nominalne varijable u SPSS-u najčešće se radi tako da se u meniju **Graph** odberu opcije **Barchart** (za štapićasti dijagram) ili **Piechart** (za pitasti dijagram ili "burek"). Za crtanje štapićastog dijagrama, u prozoru za dijalog **Bar Charts** treba odabrati dijagram **Simple**, a u okviru **Data in Chart Are** uključiti opciju **Summaries for groups of cases**. Zatim se klikne na dugme **Define** i u polje **Category axes** prebaci iz levog okvira ime varijable koja se grafički prikazuje.U okviru **Bars represent** treba odabrati (u zavisnosti od potrebe) šta će predstavljati stubići na dijagramu. Ako se odabere opcija **N of cases** onda stubovi predstavljaju broj slučajeva u kategoriji. Opcija **% of cases** znači da će stubići predstavljati procente slučajeva po kategorijama a opcije **Cum. N of cases** i **Cum.**

% of cases određuju da će stubovi prikazati kumulativne učestalosti ili kumulativne relativne frekvence u procentima. Klikom na dugme **Titles** moguće je ući u poddijalog **Titles** i tamo upisati naslov (title) i podnaslov (subtitle) grafika. Klikom na dugme **Options...** može se otići u istoimeni poddijalog i definisati da li se želi grafički prikaz nedostajućih vrednosti (missing values) kao posebne kategorije. Ukoliko želimo da nedostajuće vrednosti budu prikazane kao posebna kategorija uključićemo, u poddijalogu **Options** opciju **Display groups defined by missing values**. Klikom na dugme **Continue** poddijaloga vraćamo se u osnovni prozor za dijalog **Bar Charts** i klikom na dugme **OK** (ili Enter na tastaturi) završavamo definiciju operacije. U starijim verzijama SPSS-a štapićasti dijagram će biti prikazan u prozoru **Chart Carousel** u kojem možemo (klikom na dugme **Edit** unutar ovog prozora) pozvati editor za doterivanje grafika. U novijim verzijama dijagram će biti prikazan u prozoru za ispis, a za njegovo editovanje potrebno je kliknuti dva puta mišem na sam grafik.

Primer 1: Grafički prikaz rukosti štapićastim dijagramom (U Editoru grafike SPSS-a odabiranjem opcije **Bar label style / Framed** u meniju **Attributes** postignuto je to da na stubićima piše procenat slučajeva koji pripada datoj kategoriji)

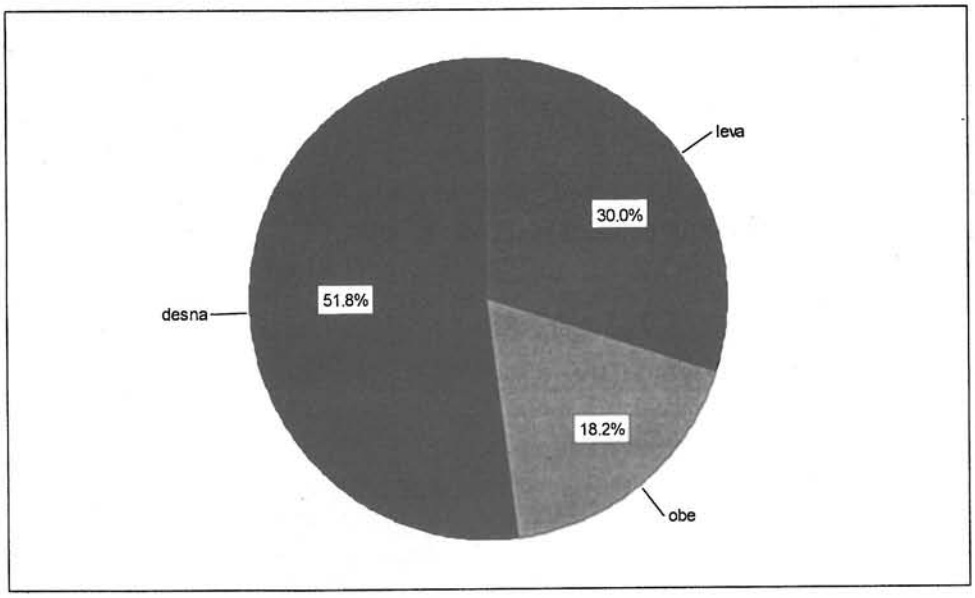


Pitasti dijagram (Piecharts)

Za prikaz kategoričke varijable pitastim dijagramom (pošto smo odabrali **Graphs/Pie)** u prozoru za dijalog **Pie Charts** treba u okviru **Data in Chart Are** uključiti (ako već nije uključena) opciju **Summaries for groups of cases** i zatim kliknuti na dugme **Define**. U prozoru za dijalog **Define Pie: Summaries for Groups of Cases** u okviru **Slices Represent** uključiti odgovarajuću opciju (u zavisnoti od toga šta želimo da odsečci, tj.kriške ovog dijagrama predstave), a u polje **Define slices by** ubaciti ime kategoričke varijable koju predstavljamo ovim dijagramom.

(Odsečak koji od celog kruga otpada na datu kategoriju kod pitastog dijagrama definisana je kao (f_k * 360)/n gde je f_k frekvenca za k-tu kategoriju a n zbir frekvenci u svim kategorijama).

Primer 2: Grafički prikaz rukosti pitastim dijagramom. (U Editoru grafike SPSS-a odabiranjem opcije **Options** u meniju **Charts**, uključivanjem **Percents** u okviru **Labels** i potom klikom na **Formats** i odabirom u polju **Position** varijante **Numbers inside text outside** postignuto je to da na "kriškama" pitastog dijagrama piše procenat slučajeva koji pripada datoj kategoriji).



****

VI. STATISTIČKO ZAKLJUČIVANJE

Dva glavna cilja statističkog zaključivanja jesu:[35]

1. **Ocenjivanje parametara**, tj. numeričkih obeležja populacije na osnovu statistika, tj. numeričkih karakteristika **slučajnog reprezentativnog** uzorka iz date populacije[36].

2. **Testiranje statističkih hipoteza**, tj. donošenje odluke (na osnovu rezultata dobijenih na uzorku) o odbacivanju ili neodbacivanju određene **statističke hipoteze** na osnovu podataka dobijenih na **slučajnom reprezentativnom** uzorku. Statistička hipoteza je iskaz o određenoj karakteristici populacije i najčešće se formuliše u obliku tzv. **nulte hipoteze.**

Pri svemu tome, treba imati na umu da statističko zaključivanje o populaciji na osnovu opservacija, tj. rezultata posmatranja dobijenih na uzorku nije moguće sa izvesnošću deduktivne logike! Naime, u deduktivnoj logici zaključak **nužno** sledi iz premisa. U statistici, međutim, zaključci o populaciji do kojih dolazimo na osnovu rezultata dobijenih na uzorku su samo **manje ili više verovatni**.

VI.1.Statističko ocenjivanje parametara

Tačkasto ocenjivanje parametara na osnovu statistika

Za psihologe, posebno one koji rade "u praksi", poznavanje teorije i postupka statističkog ocenjivanja parametara je neobično važno. Bez razumevanja ove teorije i postupaka njene primene nije moguće, između ostalog, korektno pravljenje **normi** za psihološke merne instrumente, pa ni ispravno korišćenje ovih normi u praktičnom radu pri testiranju i procenjivanju ljudi.[37]

Teorijska pozadina klasičnog pristupa ocenjivanju parametara

Na osnovu rezultata merenja neke varijable na **slučajnom reprezentativnom** uzorku možemo izračunati statističku meru uzorka, **statistik**. Dati **statistik** nam može poslužiti kao **tačkasta ocena** parametra. Ocena parametra ili statistik je, prema tome, funkcija definisana na uzorku: $G = f(X_1, X_2, ..., X_n)$. Ali, pre nego što izvedemo istraživanje i izračunamo neki

---

[35]Razumevanje postupaka statističkog zaključivanja podrazumeva poznavanje osnovnih principa teorije verovatnoće koji su veoma sažeto prikazani u **Dodatku 2** ovog priručnika.

[36] U okviru statističkog ocenjivanja razvijeno je nekoliko pristupa: klasični, Bajesovski, teorija minimaksnih ocena, teorija ekvivarijantnih ocena i sl. U ovom tekstu će biti korišćen klasična teorija ocenjivanja koju su razvili R. Fišer, H. Kramer, S.R. Rao i drugi. Podrobniji prikaz teorije ocenjivanja može se pročitati u Vuković (1997), str. 303 - 366 i Rice (1995).

[37] Norme za psihološke testove su određene numeričke vrednosti za osobinu koja se meri testom, a koje su reprezentativne za određenu grupu i u odnosu na koje se određuje izraženost date osobine kod određenog ispitanika za koga je grupa na kojoj su norme dobijene referentna grupa. Tako se, npr. inteligencija neke osobe određuje upoređenjem sa normom datog testa inteligencije za populaciju kojoj ispitanik pripada. Podrobnije o normama i njihovom korišćenju u psihologiji može se pročitati u Momirović, Wolf i Popović (1999). Statistički posmatrano, norma je zapravo ocena parametra populacije kojoj ispitanik pripada na osnovu tzv. normativnog uzorka, tj. uzorka koji će poslužiti za pravljenje normi.

statistik na osnovu određene realizacije prostog slučajnog uzorka (x_1, x_2, ...,x_n), svaka opservacija X_1, X_2, ...,X_n se teorijski posmatra kao **slučajna varijabla koja ima istu raspodelu kao i slučajna varijabla X** čije parametre treba da ocenimo.[38] Budući da pretpostavljamo da je reč o prostom slučajnom uzorku, slučajne varijable (tj. opservacije) X_1, X_2, ...,X_n su nezavisne. Kako je statistik G funkcija ovih slučajnih varijabli onda je i on **slučajna varijabla**. Distribucija verovatnoća ili funkcija gustine statistika (estimatora) je distribucija uzorkovanja (sampling distribucija). **Distribucija uzorkovanja pokazuje funkcionalnu vezu mogućih vrednosti statistika i verovatnoće (ili gustine verovatnoće) koja odgovara datoj vrednosti (ili intervalu vrednosti) statistika za sve moguće uzorke određene veličine iz date populacije.** Budući da je statistik slučajna varijabla i da ima određenu distribuciju, moguće je odrediti i očekivanu vrednost i varijansu statistika na isti način kao i za sve slučajne varijable. **Standardna devijacija distribucije uzorkovanja nekog statistika je standardna greška tog statistika.**

Prema **centralnoj graničnoj teoremi,** distribucije uzorkovanja mnogih statistika (M, Md_n) su aproksimativno normalne za uzorak dovoljne veličine bez obzira na oblik distribucije date varijable u populaciji. To, ipak, ne važi za sve statistike. Osim toga, koliko veliki uzorak je potreban za primenljivost centralne granične teoreme zavisi od toga o kojem statistiku je reč i kakva je distribucija date varijable u populaciji. Tako je, na primer, za S^2 potreban mnogo veći uzorak no za M. Tačne distribucije uzorkovanja pojedinih statistika u slučajevima kada centralna granična teorema nije primenljiva izvode se matematički, ili, ako to nije moguće, određuju se aproksimativno primenom tzv. **bootstrap** (čitaj: butstrep) postupka. Pojednostavljeno rečeno, bootstrap postupak se sastoji u tome da se kompjuterski simulira izvlačenje velikog broja slučajnih uzoraka iz populacije sa određenim pretpostavljenim parametrom i pravi distribucija učestalosti vrednosti koje dati statistik uzima. Ova distribucija se zatim koristi u aproksimaciji distribucije uzorkovanja datog statistika.[39]

Kada izvedemo istraživanje, dešava se **određena realizacija** prostog slučajnog uzorka. Na osnovu vrednosti opservacija dobijenih na tom uzorku dobijamo određenu vrednost statistika (estimatora), tj. određenu ocenu parametra. Prethodno navedeni teorijski principi omogućuju nam sve dalje operacije u vezi sa postupkom ocenjivanja parametara na osnovu dobijene, realizovane vrednosti statistika.

Greška uzorka

Greška uzorka je razlika između populacijske vrednosti (parametra) i uzoračke vrednosti statistika kojom ocenjujemo parametar. Tako, ako μ predstavlja aritmetičku sredinu neke varijable u populaciji, a M_i njenu ocenu na osnovu i-tog slučajnog uzorka veličine n, tada razlika $g_i = \mu - M_i$ predstavlja grešku i-tog uzorka. Naravno, kako μ praktično nikada nije poznato, to nismo u stanju ni da odredimo uzoračku grešku za neki određeni, konkretni uzorak. Međutim, smisleni zaključci se mogu donositi o verovatnoći pojave uzoračke greške određene veličine. Ovde nam je od velike pomoći Čejbiševljeva nejednakost jer na osnovu nje možemo da odredimo verovatnoću da se neki statistik razlikuje od svog očekivanja, tj. od parametra za određeni iznos. [40]

---

[38] O slučajnim varijablama se može pročitati u Dodatku 2 Priručnika. Podrobniji prikaz teorije slučajnih varijabli dat je u B. Ivanović (1966), str. 15 - 33.

[39] Preciznija formulacija i dokaz jednog oblika centralne granične teoreme, kao i bootstrap postupka može se pročitati u Rice (1995).

[40] Nejednakost Čejbiševa: $P(|G - \theta| < b) \geq 1 - \sigma^2 / b^2$

Standardna greška

Standardna greška nekog statistika je **standardna devijacija** njegove **distribucije uzorkovanja**.

Određivanje standardne greške ilustrovaćemo na primeru standardne greške za aritmetičku sredinu:

a) Ako je populacija beskonačna ili se uzorkovanje vrši sa vraćanjem, standardna greška za aritmetičku sredinu je :

$$\sigma(M) = \frac{\sigma}{\sqrt{n}}$$

pri čemu je σ standardna devijacija populacije, a n veličina uzorka.

b) Ako je reč o konačnoj populaciji veličine N, ili se uzorkovanje odvija bez vraćanja standardna greška za aritmetičku sredinu je:

$$\sigma(M) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Kada je n/N < 0.05, tj. kada je uzorak veoma mala **frakcija** populacije, izraz $\sqrt{\frac{N-n}{N-1}}$ u prethodnom obrascu se može zanemariti budući da je u tom slučaju praktično jednak jedinici.

Uočimo da je i standardna greška parametar koji ocenjujemo na osnovu statistika, tj. standardne greške dobijene na uzorku. Tako bi ocena standardne greške za aritmetičku sredinu bila za beskonačnu populaciju ili pri uzorkovanju sa vraćanjem:

$$SE(M) = \frac{S}{\sqrt{n}}$$

a za konačnu populaciju ili u slučaju kada uzorak čini frakciju populacije veću od 5% (n/N >0.05):

$$SE(M) = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Verovatnoća da će slučajna varijabla G da se po apsolutnoj vrednosti razlikuje od svog očekivanja, tj. aritmetičke sredine za b i više jedinica (b > 0) uvek je manja od količnika σ^2 / b^2. Odatle sledi da verovatnoća da standardizovana slučajna varijabla ima apsolutnu veličinu (vrednost) veću ili jednaku pozitivnom broju k je uvek manja ili jednaka $1/k^2$: $P\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}$

Karakteristike ocene (estimatora) parametra

Različiti statistici sa različitim uspehom ocenjuju parametar. Dakle, neki od njih imaju poželjnije osobine u pogledu ocene parametra od drugih.[41]
Sledeće karakteristike ocene parametra služe kao kriterijumi preferencije jedne metode procene nad drugom:

1. Nepristrasnost
2. Konzistentnost
3. Efikasnost
4. Dovoljnost

1. Za neki metod ocene parametra kažemo da daje nepristrasnu ocenu kada se sredina velikog broja ocena dobijenih na osnovu različitih uzoraka, sa povećavanjem broja tih izvučenih uzoraka, približava populacionoj vrednosti kao granici. Pojam nepristrasnosti procene se može izraziti i kao odsustvo sistematske tendencije da statistik potceni ili preceni parametar. Preciznije rečeno, statistik G je nepristrasna ocena parametra ukoliko je očekivana vrednost statistika, E(G) jednaka parametru θ čijoj oceni je statistik namenjen :

$$E(G) = \theta$$

Neki statistici mogu biti nepristrasne ocene parametra, dok drugi to nisu. Može se dokazati da je očekivana vrednost varijanse definisane kao:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - M)^2}{n}$$

jednaka σ^2 [(n-1) /n)], tj. manja je od σ^2. Da bi varijansa uzorka bila nepristrasna ocena varijanse populacije potrebno ju je računati po sledećem obrascu:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - M)^2}{n-1}$$

Aritmetička sredina kao statistik, ako se računa na standardni način, daje nepristrasnu ocenu parametra jer je:

$$E(M) = E\left(\frac{\sum_{1}^{n} x_i}{n}\right) = \frac{1}{n}\sum_{1}^{n} E(x_i) = \frac{1}{n}\sum \mu = \frac{1}{n} n\mu = \mu .$$

[41] Budući da je u statističkoj literaturi na našem jeziku to uobičajeno, u ovom tekstu se terminom ocena označavaju i ocenjivač (estimator) i ocena (estimate) parametra. Tako je, npr. aritmetička sredina uzorka ocenjivač (estimator) parametra (aritmetičke sredine populacije), a neka konkretna realizacija, tj određena vrednost aritmetičke sredine uzorka je ocena parametra. Pretpostavljamo da će iz konteksta čitaocu biti jasno koja od ovih značenja termin ocena ima u datom kontekstu.

Da bi standardna devijacija bila nepristrasna ocena populacijske standardne devijacije trebalo bi je računati po obrascu:

$$S' = [1+1/(4n-4)]*S$$

pri čemu je S koren iz varijanse koja je nepristrasna, tj. koja je dobijena po obrascu sa n-1 u imeniocu.

2. Statistik daje konzistentnu ocenu parametra kada se verovatnoća njegovog približavanja parametru bliži 1 sa povećanjem veličine uzorka:

$$n \to \infty \Rightarrow \quad P(|G - \theta| < b) \to 1$$

Aritmetička sredina kao statistik je konzistentna ocena parametra jer na osnovu Čejbiševljeve nejednakosti za nju sledi:

$$P(|M - \mu| < b) \geq 1 - \sigma^2 / n\, b^2$$

Sa povećanjem n, $\sigma^2 / n\, b^2 \to 0$, pa $P(|M - \mu| < b) \to 1$

3. Efikasnost ocene je količnik minimalne srednje kvadratne greške i srednje kvadratne greške statistika G koji služi kao ocena parametra:

$$Ef.(G) = E(G_0 - \theta)^2 / E(G - \theta)^2$$

U gornjem izrazu G_0 je ocena koja ima minimalnu srednju kvadratnu grešku, a srednja kvadratna greška je očekivana vrednost kvadriranih odstupanja ocene od parametra. Ako je statistik nepristrasan **srednja kvadratna greška jednaka je varijansi distribucije uzorkovanja tog statistika**. Ako je **Ef.(G) = 1**, onda je G **efikasna ocena** parametra θ.

4. Statistik je dovoljna ocena parametra kada sadrži sve informacije koje su na raspolaganju u podacima. Na primer, medijana izračunata iz intervalnih podataka ne koristi informacije o veličini razlike među rezultatima, dok aritmetička sredina koristi sve informacije sadržane u takvim podacima. U ovom slučaju aritmetička sredina je dovoljna ocena, a medijana nije.

Metode ocenjivanja

Postoji veliki broj metoda ocenjivanja: metod momenata, Monte Carlo metod, metod najmanjih kvadrata, metod maksimalne verodostojnosti i drugi.

Definisaćemo ukratko samo metod maksimalne verodostojnosti jer se on sve češće koristi.[42]

[42] Potpunije objašnjenje metoda maksimalne verodostojnosti može se pročitati u Rice (1995), str.253 - 272 ili Ivanović (1966), str.281 - 284.

Metod maksimalne verodostojnosti (maximum likelihood)

Metod se sastoji u izboru one ocene (kao vrednosti ocenjivanog parametra) koja maksimizuje funkciju L $[(x_1, x_2, ...,x_n) \mid \theta]$. Dakle, biramo kao ocenu parametra onu **vrednost koja rezultate dobijene na slučajnom uzorku čini najverovatnijim za datu vrednost parametra**. Ocena aritmetičke sredine populacije dobijena ovom metodom je

$$M = \frac{\sum_{i=1}^{n} x_i}{n}$$

a ocena varijanse populacije je

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - M)^2}{n}$$

Intervalno ocenjivanje parametara -Intervali poverenja

Intervale poverenja pravimo da bismo odredili granice intervala koji verovatno obuhvata parametar.
Principe konstrukcije intervala poverenja objasnićemo na primeru pravljenja intervala poverenja za aritmetičku sredinu:

1. Definišemo statistik

$$t = (M-\mu)/SE_M$$

pri čemu je M aritmetička sredina uzorka, μ aritmetička sredina date varijable u populaciji, a SE_M ocena standardne greške za aritmetičku sredinu;

2. Poznato je da statistik t ima Studentovu distribuciju uzorkovanja sa n-1 stepeni slobode;

3. Na osnovu prethodnog, odredićemo vrednost t-statistika (t_p) tako da je:

$$P(-t_p \leq t \leq t_p) = p.$$

Dakle, budući da je t slučajna varijabla sa Studentovom distribucijom uzorkovanja, možemo kao t_p odabrati vrednost za koju vredi da je verovatnoća da t bude u intervalu od $- t_p$ do $+ t_p$ jednaka željenoj verovatnoći (p). Koliko p ćemo odabrati zavisi od toga kakav interval poverenja hoćemo. Uobičajeno je da se konstruišu 99% ili 95% intervali poverenja tako da je p u prvom slučaju 0.99, a u drugom 0.95.

4. Umesto promenljive t u izrazu $P(-t_p \leq t \leq t_p) = p$ stavićemo $(M-\mu)/SE_M$ (na osnovu tačke 1.). Rešavanjem ovoga dobijamo interval poverenja:

$$P(M- t_p * SE_M \leq \mu \leq M+ SE_M t_p) = p.$$

Na osnovu dobijenog intervala poverenja očekujemo da interval koji dobijemo uključi parametar za određeni procenat (ako je p=0.95, za 95%) slučajnih uzoraka za koje je interval definisan na dati način. Istovremeno, p nam govori o uverenosti (ne pravoj verovatnoći!) da interval određen na ovaj način obuhvati parametar.

****

U programu SPSS tačkasto ocenjivanje aritmetičke sredine i varijanse populacije izvodi se po principu maksimalne verodostojnosti (maximum likelihood) tako da je varijansa ocenjena statistikom S^2 koji je dobijen po obrascu u kojem je imeniocu n, a ne n -1.
Tačkaste ocene aritmetičke sredine i standardne devijacije populacije mogu se dobiti u procedurama **Descriptives** i **Frequencies** (koje smo već objasnili u glavi **IV**) dok je u proceduri **Explore** moguće dobiti pored tačkastih i intervalne ocene aritmetičke sredine populacije (pravljenjem intervala poverenja).
Da bi se u proceduri **Explore** dobile tačkaste i intervalne ocene aritmetičke sredine populacije za određenu varijablu potrebno je, posle odabira **Statistics/Summarize/Explore** u okvir **Dependent list**, ubaciti ime te varijable. U okviru **Display** zatim uključujemo opciju **Statistics** i kliknemo na dugme **Statistics**. U prozoru poddijaloga **Explore: Statistics** uključujemo (ako treba, jer je to već uključeno po default-u!) opciju **Descriptives** a u ulazno polje **Confidence Interval for Mean** upisujemo željeni broj (obično 95 ili 99 jer najčešće pravimo 95% i 99% intervale poverenja).

Primer ispisa iz ove procedure:

```
DRUS      DRUSTVENOST

 Valid cases:        960.0   Missing cases:         .0   Percent missing:       .0


 Mean        18.4438  Std Err       .1969  Min          1.0000  Skewness     -.4983
 Median      19.0000  Variance    37.2148  Max         30.0000  S E Skew      .0789
 5% Trim     18.6597  Std Dev      6.1004  Range       29.0000  Kurtosis     -.3940
 99% CI for Mean (17.9356, 18.9519)       IQR          9.0000  S E Kurt      .1577
```

U ovom primeru na osnovu podataka na slučajnom reprezentativnom uzorku studenata Beogradskog univerziteta o varijabli društvenosti, crti ličnosti koja je merena Gilford-Zimermanovim pregledom temperamenta, dobili smo, između ostalih statistika, ocene aritmetičke sredine (**Mean**) i varijanse (**Variance**) populacije studenata BU, kao i standardnu grešku (**Std Err**) za aritmetičku sredinu.
Na osnovu 99% intervala poverenja za aritmetičku sredinu (**99% CI for Mean**) mogli bismo, grubo rečeno, zaključiti da bi u 99% slučajeva ponovljenih merenja aritmetička sredina populacije bila obuhvaćena intervalom od 17.94 do 18.95. Ispravnije rečeno, 99% intervala poverenja koje bismo napravili na ovaj način obuhvatilo bi parametar. Iz širine intervala poverenja (koji je veoma uzan) i iz standardne greške za aritmetičku sredinu (koja je veoma mala u odnosu na standardnu devijaciju) vidimo da u ovom slučaju možemo imati poverenja u aritmetičku sredinu uzorka kao ocenu parametra.
****

VI.2.Testiranje statističkih hipoteza

Forma logičkog argumenta koja je u pozadini statističkog testiranja hipoteza liči na Modus tollens ali je on ovde dat u nešto modifikovanom obliku u odnosu na deduktivnu logiku:

$A \Rightarrow B$	(Ako je A tačno onda verovatno B)
$\neg B$	(Nije B, tj. malo je verovatno da je B)
—	
$\neg A$	(A verovatno nije tačno)

Naime, empirijsku proveru određene teorije moguće je uraditi samo na osnovu ispitivanja empirijskih posledica dedukovanih na određeni način iz teorije. Na osnovu teorije koju proveravamo formulišemo istraživačku hipotezu o "stanju stvari" u realnosti. Da bismo statistički testirali koliko je empirijsko "stanje stvari" u skladu sa predviđanjima date teorije, na osnovu istraživačke hipoteze formulišemo statističku hipotezu, najčešće tzv. **nultu** hipotezu. Ona najčešće tvrdi nešto što je suprotno istraživačkoj hipotezi. U logičkom argumentu statistička hipoteza je A. Zatim, po prikupljenim podacima (u logičkom argumentu oni su predstavljeni slovom B) proveravamo koliko je opravdano da verujemo u tačnost onoga što tvrdi statistička hipoteza. Ako su dobijeni podaci vrlo malo verovatni pod pretpostavkom da je nulta, tj.statistička hipoteza tačna (dakle ako $\neg B$) onda možemo verovati da nulta, tj. statistička hipoteza nije tačna (dakle $\neg A$) i tada odbacujemo statističku hipotezu. Ipak, mada ovaj logički argument odslikava dobro način na koji se, na osnovu podataka dobijenih na uzorku, proverava opravdanost verovanja u tačnost statističke hipoteze važno je imati na umu da ovo zaključivanje nije u potpunosti zasnovano na izvesnosti deduktivne logike. Iskaz $\neg B$ govori samo o tome da je, u svetlu dobijenih podataka, **malo verovatno** da je nulta hipoteza tačna, a ne da ona sigurno nije tačna. Takođe, **verovatnoća tačnosti nulte hipoteze ovde je stvar stepena naše uverenosti u njenu tačnost, a ne precizno date verovatnoće da je statistička hipoteza tačna.**
Statistička procedura koja vodi odluci između nulte (npr. dobijena razlika izmedju dve aritmetičke sredine se može pripisati grešci uzorkovanja) i alternativne hipoteze (npr. dobijena razlika izmedju dve aritmetičke sredine se može pripisati uticaju eksperimentalnog faktora), zove se **test značajnosti.**

Statističke hipoteze najčešće tvrde nešto o parametrima ili svojstvima populacije. Statistička hipoteza koja definiše jednu vrednost parametra je prosta, a ona koja definiše više vrednosti parametara je kompozitna statistička hipoteza.Uobičajeno je da se definišu tzv. nulta hipoteza i alternativna hipoteza. Obično (mada to nije nužno) nulta hipoteza tvrdi da neke razlike ili povezanosti u populaciji nema, tj. da je ona jednaka nuli.
Neka su npr. M_1 i M_2 aritmetičke sredine za eksperimentalnu i kontrolnu grupu. Te aritmetičke sredine su ocene populacionih aritmetičkih sredina μ_1 i μ_2. Kako su obe vrednosti statistici na njih utiče greška uzorkovanja. Nulta hipoteza se otuda može formulisati na sledeći način: ne postoji razlika izmedju populacionih sredina μ_1 i μ_2. Ta hipoteza je nulta hipoteza i može se napisati kao: H_0: $\mu_1 - \mu_2 = 0$. Nulta hipoteza, dakle, postulira da nema razlike izmedju dve populacione aritmetičke sredine. Medjutim, kako je izraz $\mu_1 - \mu_2 = 0$ jednak izrazu $\mu_1 = \mu_2$, to znači da nulta hipoteza može biti formulisana i na sledeći način: dva uzorka su izvučena iz populacija koje imaju identičnu aritmetičku sredinu. Uopšte, bez obzira

na statistik koji se koristi, nulta hipoteza najčešće postulira da ne postoji razlika izmedju populacionih parametara.

Naučna (istraživačke) i statistička hipoteza

Statistička hipoteza je obično logička posledica sadržinske, istraživačke hipoteze.

Koraci u testiranju statističkih hipoteza su:

1. Definiše se nulta hipoteza-H_0;

2. Na osnovu rezultata dobijenih na slučajnom reprezentativnom uzorku iz populacije na koju se želi primeniti zaključak izračuna se odredjeni statistik koji se zove i **test-statistik** (jer služi za testiranje statističkih hipoteza);

3. Pod pretpostavkom da je H_0 tačna odredjuje se (na osnovu distribucije uzorkovanja test statistika) verovatnoća da test statistik kao slučajna varijabla uzme (na slučaj) vrednost dobijenu na datom uzorku ili veću od te vrednosti:

$$P(|G| \geq g_{dob}|H_0 \text{ tačna}) = \alpha.$$

Pri tom se uzima u obzir (pri tzv. dvosmernom testitanju koje se najčešće koristi) apsolutna vrednost test statistika (dakle, predznak statistika se ne uzima u obzir).[43] Pojedini autori distribuciju uzorkovanja test statistika koja je dobijena pod pretpostavkom da je H_0 tačna zovu **nultom distribucijom test - statistika.**

Zapamtite:Nulta distribucija uzorkovanja test-statistika **nije nužno ista** kao i distribucija uzorkovanja test-statistika **ako je H_0 pogrešna.** Distribucija uzorkovanja test-statistika ako je H_0 pogrešna koristi se za računanje **snage** statističkog testa (definicija snage testa data je u nastavku) i, u vezi sa tim, veličine uzorka koja je neophodna u određenom istraživanju.

4. Ako je α dovoljno malo odbacuje se H_0, a H_1 se smatra tačnom (i dobijena razlika ili povezanost se smatra statistički značajnom), a ako α nije dovoljno malo H_0 se ne odbacuje. Dovoljno malo α u većini istraživanja je najviše 0.05, a zavisno od prirode istraživanja može biti i manje, npr. najviše 0.001.

Dva tipa greške u odlučivanju o odbacivanju ili neodbacivanju H_0

Ne možemo statistički pokazati da li je H_0 tačna ili pogrešna, ali je možemo odbaciti ili ne odbaciti.U donošenju odluke u vezi sa nultom hipotezom H_0, moguće su dva tipa grešaka (Tabela 6.1):

[43] Dvosmerno i jednosmerno testiranje razlikuju se po tome kako su specifikovane nulta i alternativna hipoteza. Ako nulta i alternativna hipoteza specifikuju smer razlike onda je reč o jednosmernom testu. U slučaju jednosmernog testa verovatnoća da test statistik uzme (na slučaj) datu vrednost ili veću od nje odredjuje se uzimanjem u obzir predznaka statistika. U tom je slučaju potrebna vrednost test- statistika za odbacivanje Ho manja no pri dvosmernom testiranju.

a) greška tipa I javlja se kada se odbacuje H_0 (i prihvati alternativna hipoteza H_1), a tačna je nulta hipoteza, tj. H_0;

b) greška tipa II javlja se kada se ne odbacuje nulta hipoteza, H_0, a ona je pogrešna.

Da bi se odredila verovatnoća greške II tipa potrebno je da se odredi funkcija koja daje verovatnoću datog tipa greške za svaku vrednost parametra koja je specifikovana u alternativnoj hipotezi. Za specifikovanu vrednost parametra u alternativnoj hipotezi i za dati nivo alfa, verovatnoća greške II tipa je funkcija veličine uzorka: što je manji uzorak veća je verovatnoća ovog tipa greške. Za svaku odredjenu vrednost α, npr. 0.05 vrednost β je funkcija veličine uzorka i stvarne razlike.

Tabela 6.1: Tipovi grešaka pri odbacivanju ili neodbacivanju nulte hipoteze

	Tačna H_0	Pogrešna H_0
Odbačena H_0	Greška tipa I P (Greška tipa I) = α	Korektna odluka (Verovatnoća ove odluke je 1- β i predstavlja **snagu** statističkog testa)
H_0 nije odbačena	Korektna odluka	Greška tipa II P (Greška tipa II) = β

U opštem slučaju odlučivanje u skladu sa težnjom da se što više smanji verovatnoća greške jednog tipa povećava verovatnoću greške drugog tipa. U situacijama kada je to važno moguće je smanjiti što više verovatnoću odredjenog tipa greške. Ako npr. greška II tipa vodi tome da će kao lek biti distribuirana supstanca koja može ugroziti živote dok greška I tipa prosto znači trošenje novca nepotrebnim davanjem odredjene supstance kao leka očigledno da treba što više smanjiti verovatnoću greške II tipa. S druge strane obično je istraživačka hipoteza u skladu sa alternativnom, a ne sa nultom statističkom hipotezom, pa se nastoji na smanjenju verovatnoće greške I tipa da bi se minimizovala verovatnoća da se nekorektno tvrdi da podaci podržavaju istraživačka očekivanja.

Nivo statističke značajnosti

Jednak je α, tj. verovatnoći greške koju pravimo ako odbacujemo H_0 ako je ona tačna.

Snaga statističkog testa

P(odbacivanje H_0 | H_0 pogrešna) = 1- P(neodbacivanje H_0 | H_0 pogrešna) = 1- β.

Robustnost statističkog testa

Stepen u kojem je statistički test **osetljiv** na **neispunjenost uslova** za njegovu primenu.

VII. TESTIRANJE HIPOTEZE O NORMALNOSTI RASPODELE VARIJABLE U POPULACIJI

Mnogi statistički testovi za testiranje hipoteza pretpostavljaju da je distribucija neke varijable u populaciji normalna. Stoga je često potrebno, pre primene drugih statističkih testova proveriti da li je pretpostavka o normalnosti distribucije date varijable u populaciji zadovoljena:

H_0: Distribucija varijable u populaciji je normalna sa parametrima μ, σ^2. Slovom μ označena je očekivana vrednost (aritmetička sredina) varijable u populaciji, a znakom σ^2 varijansa varijable. Normalna funkcija gustine sa parametrima se obično piše skraćeno kao $N(\mu, \sigma^2)$.

Hi-kvadrat test normalnosti

Računamo χ^2 statistik:

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_{oj} - f_{ej})^2}{f_{ej}}$$

gde je f_{oj} empirijska, dobijena frekvenca u razredu j, f_{ej} očekivana frekvenca u razredu j na osnovu H_0, a k broj razreda.
Da li je distribucija frekvencija uzorka proizašla iz populacijske normalne distribucije?

Računamo

$$P(\chi^2 \geq \chi^2_{dobijen}) = \int_{\chi^2_{dobijen}}^{\infty} f(\chi^2)\, d\chi^2$$

tj. verovatnoću da χ^2 statistik za k - 1 - 2 stepeni slobode bude jednak ili veći od dobijenog χ^2 statistika na konkretnom uzorku. Ako je $P \leq 0.05$ odbacujemo H_0.
Test nije mnogo osetljiv na odstupanja empirijske distribucije od očekivane na osnovu pretpostavke o normalnosti. Prisutna je i arbitrarnost jer rezultati zavise od toga u koliko razreda su svrstani rezultati.

Kolmogorov - Smirnov test

Računamo statistik D, tj. najveću apsolutnu razliku empirijskih kumulativnih frekvenci i kumulativnih frekvenci koje očekujemo ako je H_0 tačno:

$$D = \max_x |F_o(x) - F_e(x)|$$

gde su F_o i F_e dobijene i očekivane (na osnovu H_0) kumulativne frekvence.

Kritičnu vrednost za D, tj. koliko najmanje treba da bude D na uzorku ($D_{dobijen}$) da bismo odbacili H_0, možemo naći preko $D_k = k(1/\sqrt{n})$.[44] D_k je kritična vrednost za D i za nivo značajnosti 0.05 k je jednako1.22, a za nivo 0.01 k je 1.52. Ako je $D_{dobijen} > D_k$ odbacujemo H_0. Ovo je najosetljiviji test normalnosti raspodele.

Standardizovani Skjunis i Kurtozis:

Standardizovani skjunis ili koeficijent zakrivljenosti dobija se kao količnik nestandardizovanog skjunisa i njegove standardne greške:

$$z_{Sk} = Sk / SE_{Sk}$$

pri čemu je

$$Sk = \frac{\sum_{i=1}^{n}(x_i - M)^3}{S * S^2} \frac{n}{(n-1)(n-2)}$$

ocena populacijskog skjunisa (videti o momentima u odeljku **IV.7.**),a

$$SE_{Sk} = \sqrt{\frac{6}{n}}$$

je standardna greška za Skjunis.

Standardizovani koeficijent izduženosti je količnik nestandardizovanog kurtozisa i njegove standardne greške:

$$z_{Ku} = Ku/SE_{Ku}$$

pri čemu je:

$$Ku = \frac{\sum_{i=1}^{n}(x_i - M)^4}{S^2 * S^2} \frac{n(n+1)}{(n-1)(n-2)(n-3)} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

ocena populacijskog kurtozisa, a

$$SE_{Sk} = \sqrt{\frac{24}{n}}$$

standardna greška za Ku.

Budući da z_{Sk} i z_{Ku} približno imaju (ako je n dovoljno veliko, tj. ako je n > 150) standardizovanu normalnu distribuciju uzorkovanja, možemo izračunati kolika je verovatnoća da se na slučajnom uzorku dobije toliki (ili veći) koeficijent zakrivljenosti ili izduženosti ako je H_0 tačno. Ako je ta verovatnoća jako mala (npr. manja od .05) odbacuje se nulta hipoteza (o normalnosti raspodele varijable u populaciji)

---

[44] Kolmogorov je pokazao da je, ako je $l = D\sqrt{n}$, pri porastu n, raspodela k(l), tj. Kolmogorovljeva funkcija: $k(l) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2l^2}$ nezavisna od toga kakva je $F_e(x)$. Prema tome, $P(l \geq l_a) = P(D\sqrt{n} \geq l_a) = 1 - k(l) = \alpha$. Isto tako, $P(D\sqrt{n} \geq l) = 1 - k(l)$.

****

Testiranje normalnosti raspodele u SPSS programu:

1. Kolmogorov -Smirnov test ako su parametri (aritmetička sredina i varijansa populacije) poznati:

U meniju **Statistics/Nonparametric Tests/1 Sample K-S** u dijalogu **One-Sample Kolmogorov-Smirnov test** u **Test variable list** uneti ime varijable čija normalnost se testira. U okviru **Test Distribution** odabrati **Normal** (to je i default opcija). Klikom na **Ok** operacija je definisana, a na ispisu se između ostalog dobija i α u koloni **2-Tailed P**. Međutim, ova verovatnoća je primenljiva u testiranju normalnosti raspodele samo ukoliko su parametri (aritmetička sredina μ i varijansa σ^2) poznati. Ukoliko se ovi parametri ocenjuju na osnovu podataka dobijenih na uzorku onda je potrebno koristiti Lilliefors verovatnoću (pogledati objašnjenje pod 1a u nastavku teksta).

1a) Ukoliko se parametri normalne funkcije gustine (aritmetička sredina i varijansa varijable u populaciji) ocenjuju na osnovu podataka dobijenih na uzorku onda je za odbacivanje (ili neodbacivanje H_0) bolje koristiti Lillieforsovu modifikaciju Kolmogorov - Smirnovljevog testa, koja se dobija u meniju **Statistics/Summarize/ Explore** odabirom (u okviru **Display** dijaloškog prozora **Explore**) opcije **Plots** i potom (posle klika na dugme **Plots**) uključivanjem opcije **Nomality plots with tests**. Pri tom, ako želimo samo test normalnosti potrebno je u okviru **Boxplots** uključiti opciju **None**, a u okviru **Descriptives** isključiti opciju **Steam-and-leaf**. Podrazumeva se da je u okvir **Dependent List**, u prozoru **Explore**, ubačeno ime varijable za koju testiramo normalnost. U okvir **Factor List** (ukoliko test radimo na celom uzorku) nije potrebno unositi ništa.

Primer ispisa iz ove procedure:

	Statistic	**df**	**Significance**
K-S (Lilliefors)	**.0665**	**120**	**> .2000**

U koloni **Statistic** data je vrednost D statistika a u koloni **Significance** verovatnoća da statistik D bude, za dati broj stepeni slobode, jednak ili veći od dobijenog ako je H_0 tačno, tj. ako je distribucija varijable u populaciji normalna. Ukoliko je ova verovatnoća manja ili jednaka 0.05 (ili manja ili jednaka nekoj drugoj vrednosti odabranoj za odbacivanje H_0) onda H_0 odbacujemo. Verovatnoća na osnovu koje odlučujemo da li ćemo odbaciti H_0 u ovoj proceduri izračunata je pod pretpostavkom da parametre nemamo već ih ocenjujemo na osnovu aritmetičke sredine i varijanse uzorka. Broj stepeni slobode (kolona **df**) jednak je veličini uzorka.

Na žalost, pojedine verzije SPSS-a pre verzije 7.5 pri određivanju verovatnoće u Lilieforsovoj modifikaciji Kolmogorovljevog testa kao statistik D ne uzimaju apsolutnu već aritmetički najveću razliku (što je pogrešno!).

Ukoliko je uzorak jednak ili manji od 50, pored ove Lillieforsove modifikacije Kolmogorov-Smirnovljevog testa biće prikazani i rezultati Šapiro-Vilksovog testa normalnosti (za male uzorke):

	Statistic	df	Significance
Shapiro-Wilks	**.9681**	**41**	**.4295**

Ukoliko je verovatnoća da je Šapiro-Viksov statistik jednak ili veći od dobijenog (kolona **Significance**) ako je H_0 tačno jednaka ili manja od .05 odbacujemo H_0. Broj stepeni slobode u ovom testu jednak je veličini uzorka.
Kod malih uzoraka bolje je, pri odlučivanju o sudbini H_0, osloniti se na Šapiro-Vilksov nego na Kolmogorov-Smirnovljev test normalnosti.

2. Hi-kvadrat test:

Potrebno je napraviti distribuciju učestalosti sa razredima pre definisanja ove operacije i imati pripremljene očekivane frekvence za svaki razred ako je H_0 tačno. Potom se u dijalogu menija **Statistics/Nonparametric Tests/Chi-Square** u okvir **Expected Values** unose (upisivanjem i potom klikom na **Add**) očekivane frekvence od najnižeg do najvišeg razreda. Očekivane verovatnoće, pod pretpostavkom da je distribucija varijable u populaciji normalna, mogu se u SPSS-u dobiti uz pomoć funkcije **CDF.normal** u meniju **Transform / Compute**. Korišćenje funkcije CDF.normal objašnjeno je u Dodatku 2. Očekivane frekvence se, zatim, dobijaju tako što se verovatnoće izračunate uz pomoć funkcije CDF.normal pomnože brojem rezultata, tj. veličinom uzorka.

3. Testiranje normalnosti pomoću standardizovanog Skjunisa i Kurtozisa:

Skjunis i kurtozis, kao i njihove standardne greške za računanje standardizovanog skjunisa i kurtozisa, mogu se dobiti u menijima **Statistics/Summarize/Frequencies** ili **Statistics/Summarize/Descriptives** izborom odgovarajućih opcija. U poddijalogu **Statistics** procedure **Frequencies** dugmad za odabir skjunisa i kurtozisa nalaze se u okviru **Distribution,** dok se u proceduri **Descriptives** ove opcije nalaze u poddijalogu **Options**. Zatim je potrebno nestandardizovani skjunis i kurtozis podeliti odgovarajućim standardnim greškama. Ukoliko je bilo koja od ovih dveju standardizovanih vrednosti veća od 1.96 (za nivo 0.05) ili 2.58 (za nivo 0.01) H_0 se može odbaciti. Naime, budući da standardizovani skjunis i kurtozis za velike uzorke imaju približno standardizovanu normalnu distribuciju uzorkovanja, verovatnoća da se na uzorku dobije standardizovani skjunis ili kurtozis veći ili jednak 1.96 (2.58) je manja od 0.05 (0.01). Prema tome, 1.96 je **kritična vrednost** za ova dva statistika za odbacivanje H_0 na nivou 0.05, a 2.58 za odbacivanje H_0 na nivou 0.01.
****

VIII. TESTIRANJE HIPOTEZE O ARITMETIČKOJ SREDINI POPULACIJE

Testiranje hipoteze da je aritmetička sredina populacije jednaka nekoj vrednosti

Pre izvođenja ovog statističkog testa aritmetička sredina slučajnog uzorka (M) je slučajna varijabla pa se na nju mogu primeniti mnoge teoreme o slučajnim varijablama od kojih je najvažnija centralna granična teorema.

Centralna granična teorema

Ako populacija ima konačnu varijansu σ^2 i aritmetičku sredinu μ, distribucija aritmetičkih sredina uzorka od n nezavisnih opservacija se približava normalnoj sa varijansom σ^2 / n i aritmetičkom sredinom μ kako n raste. Kada je n veoma veliko distribucija uzorkovanja (sampling distribution) za M je približno normalna.[45]

Nešto opštiji oblik ove teoreme glasi: Ako napravimo varijablu bilo kakvom linearnom kombinacijom većeg broja varijabli koje su nid, tj. nezavisno i identično distribuirane (ne nužno normalno) nova varijabla će težiti da ima normalnu distribuciju, pri čemu je tendencija utoliko jača ukoliko je nova varijabla napravljena od većeg broja varijabli.

1. Postavljamo nultu i alternativnu hipotezu :

$$H_0: \mu = \mu_0 \Leftrightarrow \mu - \mu_0 = 0.$$

$$H_1: \mu \neq \mu_0 \Leftrightarrow \mu - \mu_0 \neq 0.$$

U tim hipotezama μ_0 je vrednost kojoj je po nultoj pretpostavci jednaka aritmetička sredina populacije (iz koje je uzorak) na datoj varijabli.

2. Računamo statistik:

$$t = (M-\mu_0) / SE_M$$

gde je SE_M ocena standardne greške za aritmetičku sredinu, M aritmetička sredina uzorka a μ_0 pretpostavljena vrednost aritmetičke sredine u populaciji.

Statistik t ima, ako je H_0 tačna, Studentovu T- funkciju uzorkovanja sa n-1 stepeni slobode. [46]

3. Računamo $P(|t| \geq t_{dobijeno} \mid H_0 \text{ tačno}) = \alpha$. Ako je $\alpha \leq .05$ odbacujemo H_0.

****

[45] Precizna formulacija i dokaz jednog oblika centralne granične teoreme može se pročitati u Rice (1995), str. 169 - 173.

[46] T- statistik se može prikazati i kao razlomak u čijem je brojiocu varijabla koja ima standardizovanu normalnu raspodelu, a u imeniocu je kvadratni koren varijable koja ima Hi-kvadrat raspodelu sa odgovarajućim brojem stepeni slobode (u ovom slučaju n - 1) i koja je podeljena odgovarajućim brojem stepeni slobode. Takav razlomak, kao što je poznato iz matematičke statistike, ima uvek Studentovu raspodelu čiji parametar su stepeni slobode iz imenioca razlomka.

Testiranje hipoteze o aritmetičkoj sredini populacije u SPSS-u:

U okvir **Test variable** dijaloga **One-Sample t-test** menija **Statistics/Compare Means/ One-Sample t-test** treba uneti ime varijable čija se aritmetička sredina u populaciji testira. U polje **Test value** se unosi pretpostavljena vrednost aritmetičke sredine u populaciji, tj.μ_0. Između ostalog u ispisu su dati t-statistik, broj stepeni slobode (n-1) i **2-Tail Sig** tj. α

Primer: Da li uzorak pripada populaciji sa aritmetičkom sredinom 15 na varijabli društvenosti? (Pri definisanju obrade u polje **Test value** uneta je vrednost 15).

One Sample t-tests

Variable	Number of Cases	Mean	SD	SE of Mean
DRUS DRUSTVENOST	960	18.4438	6.100	.197

Test Value = 15

Mean Difference	95% CI Lower	95% CI Upper	t-value	df	2-Tail Sig
3.44	3.057	3.830	17.49	959	.000

Aritmetička sredina uzorka je 18.44 a njena standardna greška je 0.197. Pod nazivom **95% CI** date su donja i gornja granica 95% intervala poverenja za razliku između aritmetičke sredine populacije i aritmetičke sredine našeg uzorka (**Mean Difference**).
T statistik (**t-value**) je statistički značajan (vrednost u koloni **2-Tail Sig** je manja od 0.05) za 959 stepeni slobode (n-1). Dakle, verovatnoća da dobijemo t statistik po apsolutnoj vrednosti 17.49 ili veći od toga je veoma mala ako uzorak pripada populaciji sa aritmetičkom sredinom na varijabli društvenosti jednakom 15.

Zaključak: Uzorak po svoj prilici ne pripada populaciji u kojoj je μ =15 u pogledu društvenosti.
****

IX. TESTIRANJE ZNAČAJNOSTI RAZLIKA IZMEĐU ARITMETIČKIH SREDINA- t test za nezavisne i zavisne uzorke

Za nezavisne uzorke:[47]

Struktura podataka:

$E = \{ e_i ; i = 1,...,n\} \subset P = P_1 \cup P_2$ je slučajni uzorak izvučen iz populacije sastavljene od dve subpopulacije;
$Q = \{ q_k ; k = 1,2\}$ je kategorička varijabla koja definiše pripadnost ispitanika određenoj subpopulaciji;
v je kvantitativna varijabla.

$\mathbf{x_1} = E_1 \otimes v =$

1	x_{11}
2	x_{12}
.	.
n_1	x_{1n1}

$\mathbf{x_2} = E_2 \otimes v =$

1	x_{21}
2	x_{22}
.	.
n_2	x_{2n2}

Nulta i alternativna hipoteza:

H_0: $\mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0$.
H_1 : $\mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0$.

Ako su varijanse subpopulacija homogene, tj. ako je $\sigma^2_1 = \sigma^2_2$. test statistik:

$$t = [\,|M_1 - M_2| - (\mu_1 - \mu_2)] / SE_{DM}$$

ima Studentovu t -funkciju gustine sa $(n_1-1)+(n_2-1)$ stepena slobode pri čemu su n_1 i n_2 veličine uzoraka.

U izrazu za t statistik M_1-M_2 je razlika između aritmetičkih sredina uzoraka D_M , a

$$SE_{DM} = \sqrt{S^2 (1/n_1 + 1/n_2)}$$

je ocena standardne greške za razliku između aritmetičkih sredina (pri čemu je $S^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$ ocena varijanse varijable u populaciji dobijena na uzorku).

[47] O zavisnim i nezavisnim uzorcima može se pročitati u Č. Dragićević: Statistika za psihologe.

Računamo $\alpha = P(|t| \geq t_{dobijeno} \mid H_0 \text{ tačno}) = \alpha$. Ako je $\alpha \leq .05$ odbacujemo H_0.

Ako varijanse subpopulacija nisu homogene tj. ako je $\sigma^2_1 \neq \sigma^2_2$:
računamo

$$t' = [|M_1 - M_2| - (\mu_1 - \mu_2)] / \sqrt{S_1^2 / n_1 + S_2^2 / n_2} .$$

Statistik t' ima Behrens-Fisherovu distribuciju uzorkovanja. Verovatnoća za t' se može aproksimirati preko Studentove t-funkcije gustine, ali se broj stepeni slobode prema predlogu koji su dali Welch & Satterthwate određuje na sledeći način:

$$df = \frac{\left[(S_1^2 / n_1) + (S_2^2 / n_2)\right]^2}{\left[(S_1^2 / n_1)^2 / (n_1 - 1)\right] + \left[(S_2^2 / n_2)^2 / (n_2 - 1)\right]}$$

U slučaju da smo odbacili nultu hipotezu (H_0: $\mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0$) ima smisla izračunati meru intenziteta razlika između aritmetičkih sredina, tj. pointbiserijski koeficijent korelacije:

$$r_{pb} = \frac{M_1 - M_2}{S_t} * \sqrt{p_1 p_2}$$

gde su M_1 i M_2 aritmetičke sredine grupa, S_t je standardna devijacija svih rezultata a $p_1 = n_1/n$ i $p_2 = n_2/n$ (gde je $n_1 + n_2 = n$) proporcije ispitanika u svakoj grupi. Što je r_{pb} bliži jednici (bez obzira na predznak) to je razlika veća, a što je bliži nuli to je razlika manja (i često manje praktično važna).
Inače, pointbiserijski koeficijent korelacije se može izračunati ukoliko znamo vrednost t-statistika jer je $r_{pb}^2 = \frac{t^2}{t^2 + df}$ budući da je $t = r_{pb} \frac{\sqrt{n-2}}{\sqrt{1 - r_{pb}^2}}$

Jedna od često korišćenih mera intenziteta razlika između aritmetičkih sredina, tj. mera efekta nezavisne varijable (faktora) na zavisnu je tzv. kvadrirana omega:

$$\omega^2 = \frac{t^2 - 1}{t^2 + df + 1}$$

Očigledno je da će za iste podatke kvadrirana omega biti manja od koeficijenta pointbiserijske korelacije tako da već vrednost kvadrirane omege od 0.20 govori o jakom efektu.

****

U SPSS programu testiranje značajnosti razlika između aritmetičkih sredina za nezavisne uzorke radi se tako da se odabere meni **Statistics/Compare Means/Independent Samples T-test**. U dijalog prozoru **Independent Samples T-test** se kao **Test variable** unosi (dvostrukim klikom na njeno ime u levom okviru sa spiskom svih varijabli ili jednim klikom na ime varijable i potom na strelicu ispred okvira Test variable) ime kvantitativne varijable po kojoj se utvrđuje da li subpopulacije imaju jednake aritmetičke sredine a u polje **Grouping Variable** ime kategoričke varijable koja deli populaciju na dve subpopulacije. Kada se ime kategoričke varijable ubaci u polje Grouping variable treba kliknuti na dugme **Define Groups** i potom u polja **Group1** i **Group2** upisati brojeve

kojima su označene prva i druga grupa u fajlu sa podacima.[48]. Klikom na **Continue** i **OK** završeno je definisanje operacije.
Zavisno od toga da li Levenov test sugeriše opravdanost pretpostavke o homogenosti varijansi subpopulacija koristi se t-test u redu **equal** u SPSS 6.1.ili **Equal variances assumed** u SPSS 8.0. (kada je **P** u Leven-ovom testu veće od 0.05) ili **unequal** (kada je P u Levenovom testu manje ili jednako 0.05 tj. kada je pretpostavka o homogenosti varijansi subpopulacija sumnjiva).
Test statistik je dat u koloni **t values** a broj stepeni slobode u koloni **df**. U **koloni 2-Tail Sig** data je verovatnoća da test statistik kao slučajna varijabla uzme vrednost dobijenu na datom uzorku ili veću od toga bez obzira na predznak. Ukoliko je 2-Tail Sig veći od .05 H_0 se ne odbacuje a ukoliko je manji ili jednak .05 H_0 se odbacuje. (Ukoliko se odredi drugi nivo značajnosti, npr. 0.001, a ne 0.05 onda se 2-Tail Sig poredi sa vrednošću koju je odabrao istraživač).
Ukoliko nas zanima ne samo da li je razlika dveju AS statistički značajna već i (ako je značajna) koliko je izražena moguće je t-test uraditi pomoću menija **Statistics/Compare Means/Means** i videti u ispisu vrednost Eta koeficijenta korelacije između pripadnosti dvema subpopulacijama i kvantitativne varijable. Što je ovaj koeficijent bliži jedinici to je razlika izraženijeg intenziteta a što je bliže nuli to je razlika manje izražena.

Primer 1: Da li Skala uticaja događaja (Impact of event scale) dobro diskriminiše ispitanike sa kliničkom dijagnozom Postraumatskog stresnog poremećaja (PTSP) od onih koji nemaju dijagnistifikovan PTSP?
Ovo pitanje je identično pitanju: Da li se subpopulacije ljudi sa PTSD-om i onih koji nemaju PTSD razlikuju u pogledu prosečnih mera na IES skali? Budući da je reč o dvema različitim grupama ispitanika pri čemu izbor nekog člana u grupu sa PTSD-om nije uticao na izbor člana u grupu bez PTSD-a reč je o nezavisnim uzorcima. Pripadnost grupi sa ili bez PTSD-a je faktor (i to neponovljen) a "zavisna" varijabla je stepen traumatizovanosti iskazan ukupnim rezultatom na IES skali.
Ispis koji je ovde prikazan dobijen je u proceduri **Independent samples t-test**. Ova analiza je moguća i u procedurama za jednofaktorsku univarijatnu analizu varijanse koje su prikazane u jednom od narednih poglavlja.

t-tests for Independent Samples of PTSD Postraum. stresni poremecaj-klin.dija

Variable	Number of Cases	Mean	SD	SE of Mean
IES UKUPNI SKOR NA SKALI UTICAJA DOGAĐAJA				
ne	145	39.5103	21.570	1.791
da	53	50.8868	15.490	2.128

Mean Difference = -11.3764

Levene's Test for Equality of Variances: F= 9.910 P= .002

	t-test for Equality of Means				95%
Variances	t-value	df	2-Tail Sig	SE of Diff	CI for Diff
Equal	-3.52	196	.001	3.232	(-17.751, -5.002)
Unequal	-4.09	128.52	.000	2.781	(-16.880, -5.873)

[48] U okviru Define Groups postoji opcija **Cut Points** koja se koristi onda kada se grupe ne definišu kategoričkom varijablom već tako što se na nekoj kvantitativnoj varijabli odredi vrednost koja deli ispitanike u dve grupe

Pre dalje analize, treba prvo pogledati ishod Leveneovog testa jednakosti varijansi subpopulacija u pogledu zavisne varijable. Budući da su varijanse subpopulacija u pogledu zavisne varijable po svemu sudeći heterogene (F=9.91 i statistički je značajan) u ovom slučaju nije preporučljivo (pogotovu kada su uzorci nejednake veličine) koristiti ishode standardnog t-testa koji su dati u redu **Equal** u delu ispisa **t-test for Equality of Means**. Bolje je pogledati ishode koji su dobijeni modifikovanim t-testom a koji su dati u redu **Unequal**. T statistik je 4.09 a verovatnoća u koloni **2-tail Sig** sugeriše da ima smisla odbaciti H_0. Mogli bismo, dakle, zaključiti da su aritmetičke sredine subpopulacija u pogledu stepena traumatizovanosti (sa i bez PTSD-a) različite.
Za određivanje veličine ove razlike možemo (primenom procedure **Statistics/Correlate/Bivariate**) izračunati pointbiserijsku korelaciju između binarne varijable PTSD i kvantitativne varijable IES. Ona je u ovom slučaju 0.24 što govori da IES skala jeste diskriminativna ali ne u meri u kojoj bismo to očekivali od skale koja je namenjena merenju stepena traumatizovanosti ljudi kao posledicom iskustva koje prevazilazi uobičajena ljudska iskustva.

Primer2: Da li se subpopulacije ljudi sa PTSP poremećajem i onih koji nemaju PTSP razlikuju u pogledu prosečne opšte intelektualne sposobnosti? Opšta intelektualna sposobnost je procenjena rezultatom na testu S-1.

```
t-tests for Independent Samples of PTSD    Postraum. stresni poremecaj-klin.dija

                              Number
Variable                     of Cases       Mean          SD   SE of Mean
-----------------------------------------------------------------------
S_1  UKUPNI REZULTAT NA TESTU SPOSOBNOSTI S-1

ne                             75        13.4533       7.579         .875
da                             47        12.1277       7.392        1.078
-----------------------------------------------------------------------

         Mean Difference = 1.3257

         Levene's Test for Equality of Variances: F= .157   P= .692

      t-test for Equality of Means                                   95%
Variances   t-value       df    2-Tail Sig     SE of Diff          CI for Diff
-----------------------------------------------------------------------------
Equal           .95      120          .344          1.397       (-1.440, 4.091)
Unequal         .95    99.68          .342          1.389       (-1.429, 4.081)
-----------------------------------------------------------------------------
```

Očigledno je da su varijanse subpopulacija (sa i bez PTSP-a) homogene u pogledu varijabilnosti opšte intelektualne sposobnosti (Leveneov F statistik je 0.157 i nije statistički značajan jer je P= 0.692, tj.veće je od 0.05. Stoga u ovom primeru možemo zaključke doneti na osnovu standardnog t-testa (red **Equal**) u delu ispisa **t-test for Equality of Means**. T-statistik od 0.95 nije statistički značajan za 120 stepeni slobode (p=0.344). Treba uočiti i da u ovom slučaju 95% interval poverenja za razliku između aritmetičkih sredina obuhvata i vrednost 0. To je samo drugi način da utvrdimo da razlika dveju grupa nije statistički značajna, tj da se subpopulacije ne razlikuju u pogledu prosečne inteligencije. U prethodnom primeru (v. Primer 1), u kojem su se subpopulacije razlikovale po prosecima na zavisnoj varijabli, 95% intervalom poverenja za razliku između aritmetičkih sredina nije bila obuhvaćena vrednost 0.
****

Za zavisne uzorke pod modelom diferencija

Struktura podataka:

$E = \{ e_i : i = 1,....,n\} \subset P$ je slučajni uzorak izvučen iz populacije;
v je kvantitativna varijabla

$\mathbf{x_1} = E_1 \otimes$ v u trenutku t_1=

1	x_{11}
2	x_{12}
.	.
n_1	x_{1n1}

$\mathbf{x_2} = E_2 \otimes$ v u trenutku t_2 =

1	x_{21}
2	x_{22}
.	.
n_2	x_{2n2}

$\mathbf{d} = \mathbf{x_2} - \mathbf{x_1}$ je vektor čiji su elementi diferencije rezultata u paru ($d_i = x_{1i} - x_{2i}$).

Nulta i alternativna hipoteza:

H_0: $\mu_d = 0$. Aritmetička sredina razlika u populaciji jednaka je nuli.
H_1 : $\mu_d \neq 0$.

Test statistik

$$t = (M_d - \mu_d) / SE_{Md}$$

($M_d = \frac{\sum_{i=1}^{n} d_i}{n}$ je aritmetička sredina diferencija, a $SE_{M_d} = \sqrt{\left[\frac{\sum_{i=1}^{n}(d_i - M_d)^2}{n-1}\right]} * 1/n$ je standardna greška za aritmetičku serdinu razlika) ima pod H_0 Studentovu t-distribuciju sa n - 1 stepeni slobode gde je n broj diferencija.

Računamo $\alpha = P(|t| \geq t_{dobijeno} | H_0$ tačno) = α. Ako je $\alpha \leq .05$ odbacujemo H_0.
Odbacivanje H_0, zapravo, znači da opservirana razlika ne može biti objašnjena greškom uzorkovanja, već delovanjem nekog sistematskog faktora (npr. eksperimentalnog tretmana). Ako se takav rezultat dobije, onda se kaže da je on statistički značajan. Ako se verovatnoća dobijanja te razlike ne može smatrati malom, onda se ona ima smatrati posledicom greške uzorkovanja. Tada se ne može (dovoljno) pouzdano zaključiti da su razlike posledica eksperimentalnog tretmana.

****

U SPSS programu testiranje značajnosti razlika između aritmetičkih sredina za zavisne uzorke radi se tako da se odabere meni **Statistics/Compare Means/Paired Samples T-test**. U dijalog prozoru **Paired Samples T-test** treba kliknuti sukcesivno na imena obe

varijable u kojima su podaci za poređenje i potom na strelicu ispred okvira **Paired Variables list**. U okviru Paired Variables list pojaviće se imena obeju varijabli razdvojena znakom - tj. na taj način je definisan vektor razlika dveju varijabli. Klikom na dugme **OK** završeno je definisanje operacije.
U ispisu je dat broj parova rezultata (**Number of pairs**), korelacija dva niza podataka i njena značajnost (kolone **Corr** i **2-Tail Sig**), Aritmetičke sredine (**Means**),standardne devijacije (**SD**) i standardne greške (**SE of Mean**) obeju varijabli a u odeljku **Paired Differences** aritmetička sredina diferencija (Mean), standardna devijacija (SD) i standardna greška aritmetičke sredine diferencija (SE of Mean).U koloni **2-Tail Sig** data je verovatnoća da za n-1 (n je broj parova rezultata) stepeni slobode (**df**) t-statistik (**t-value**) uzme dobijenu vrednost ili veću od nje ako je H_0 tačno.
Ukoliko je 2-Tail Sig veći od dozvoljene verovatnoće greške za odbacivanje tačne H_0 (najčešće je to .05) H_0 se ne odbacuje a ukoliko je manji ili jednak toj vrednosti H_0 se odbacuje. Ukoliko je korelacija dveju varijabli (kolona **Corr**) blizu nule ili jednaka nuli nema smisla testirati kvantitativne razlike jer postoje kvalitativne.

Primer: Da li stres ima efekat na uspešnost u obavljanju određene aktivnosti? Ista grupa obavljala je određenu aktivnost bez prisustva stresa i sa prisutnim stresom. Redosled izlaganja ovim dvema eksperimentalnim situacijama je balansiran unutar grupe tako da je polovina ispitanika bila prvo pod stresom a potom bez stresa a druga polovina ispitanika prvo bez a zatim sa stresom. Budući da poredimo **isti** uzorak u pogledu uspešnosti u dvema eksperimentalnim situacijama reč je o zavisnim uzorcima. Faktor stresa je, dakle, u ovom slučaju ponovljen. Zavisna varijabla je uspešnost (u situaciji bez stresa-BS a u situaciji sa stresom-S).

t-tests for Paired Samples

Variable	Number of pairs	Corr	2-tail Sig	Mean	SD	SE of Mean
S				8.5000	.972	.307
	10	.355	.314			
BS				10.1000	1.449	.458

Paired Differences			t-value	df	2-tail Sig
Mean	SD	SE of Mean			
-1.6000	1.430	.452	-3.54	9	.006
95% CI (-2.623, -.577)					

Vidimo da je razlika između proseka uspešnosti -1.60 = 8.50 - 10.10. U koloni **corr** vidimo da je korelacija između rezultata u dvema situacijama 0.36. T-statistik je 3.54 (predznak ovog statistika trenutno nije bitan, jer vršimo dvosmerno testiranje. Verovatnoća da dobijemo takav ili veći t statistik ako stres nema efekta (tj. ako je H_0 tačna) je vrlo mala (0.006). Zato odbacujemo H_0 i zaključujemo da stres ima efekta na stepen uspešnosti u obavljanju date aktivnosti. Treba uočiti da u ovom slučaju nemamo posla sa tzv. realno postojećim već eksperimentalnim populacijama (ljudi u situaciji bez stresa i u situaciji sa stresom). Uočimo isto tako da granice 95% intervala poverenja za aritmetičku sredinu razlika (**95% CI**) ne obuhvataju nulu. To takođe govori o potrebi odbacivanja H_0.
****

X. ANALIZA VARIJANSE

U ovom delu teksta analiza varijanse tretirana je prevashodno na "tradicionalan" način, tj. onako kako je ona prikazana u mnogim udžbenicima koji su posvećeni statističkoj analizi eksperimentalnih podataka. U odeljku o multiploj regresionoj analizi jednofaktorijalna univarijatna anliza varijanse prikazana je kao specijalni slučaj multiple regresione analize.[49] Zapravo, obe metode se u novije vreme tretiraju zajedno pod imenom **generalni linearni model.**[50]

Analiza varijanse je metoda koja omogućuje deljenje varijanse dobijene u rezultatima istraživanja (eksperimenta) na delove, pri čemu je moguće te delove varijanse dovesti u vezu sa nekim poznatim izvorom, uzrokom ili faktorom. Ovom metodom je, dakle, moguće proceniti veličinu varijanse koju objašnjava neki faktor, i proceniti da li je taj deo varijanse veći od očekivanja na osnovu nulte hipoteze. Metodu je predložio jedan od najpoznatijih statističara Sir Ronald Aylmer Fisher 1923.godine, a način na koji je Fišer postavio i rešavao ovaj problem dat je u posebnom odeljku ovog teksta.

Analiza varijanse se, kao što smo već pomenuli, može tretirati i kao poseban slučaj multiple regresione analize u kojoj kao prediktori služe binarne varijable, koje definišu da li ispitanik pripada ili ne pripada datoj subpopulaciji. Isto tako, jednofaktorijalna univarijatna analiza varijanse je poseban slučaj diskriminacione analize (o kojoj će tek biti reči) kada se skup kvantitativnih varijabli na osnovu kojeg želimo da predvidimo kategoričku varijablu sastoji samo od jedne varijable.

X.1. Jednofaktorijalna univarijatna analiza varijanse

Najjednostavniji oblik analize varijanse je jednofaktorijalna (jednosmerna) univarijatna analiza varijanse. Ona je namenjena testiranju postojanja razlika između aritmetičkih sredina različitih subpopulacija (definisanih kategorijama nominalne varijable) na nekoj kvantitativnoj varijabli.

Struktura podataka:

$E = \{e_i; i = 1,\ldots,n\} \subset P = \bigcup_k^g P_k$ je slučajni uzorak izvučen iz populacije koja čini uniju g subpopulacija $Q = \{ q_k ; k = 1,2,\ldots,g\}$ je kategorička varijabla koja se sastoji od g kategorija. Ova se varijabla smatra nezavisnom i zove se još i **faktor**. Kao i svaka

---

[49] Pridev *faktorijalna (u kojem je umesto zvezdice u konkretnom slučaju brojčana odrednica: jedno-, dvo-) odnosi se na broj nezavisnih, kategoričkih varijabli, tj. faktora, dok pridev *varijatna (u kojem je umesto zvezdice u datom slučaju odrednica uni-, multi-) govori o broju zavisnih, kvantitativnih varijabli u analizi varijanse.

[50] Generalni linearni model ima oblik: $y_i = b_0 + b_1x_{1i} + b_2x_{2i} +\ldots+ b_{mi}x_{mi} + g_i$ ili, u matričnoj notaciji: $y = Xb + g$, pri čemu je y vektor rezultata u zavisnoj varijabli, X je matrica čiji su kolonski vektori prediktorske ili nezavisne varijable, b je vektor pondera, a g je vektor reziduala ili grešaka. Kao što se može videti u odeljku XIV multipli regresioni model je identičnog oblika. Analiza varijanse ima oblik: : $y = X\tau + g$, pri čemu je y vektor sa rezultatima na zavisnoj varijabli, X je tzv. matrica dizajna koju čine binarizovane nezavisne varijable, g je vektor grešaka, a elementi vektora τ su parametri za pojedine efekte. Elementi vektora τ su ukupna aritmetička sredina koja odgovara interceptu u regresionom modelu (b_o) i parametri pojedinih efekata (zavisno od nacrta analize varijanse), koji odgovaraju regresionim koeficijentima pojedinih prediktora. Dakle, reč je o suštinski istom modelu. Jedina razlika je u tome što su u analizi varijanse nezavisne varijable binarne dok su u regresionoj analizi one najčešće (mada ne nužno) kvantitativne. Podrobnija objašnjenja veze analize varijanse i generalnog linearnog modela mogu se pročitati u Everitt, 1996. i Howell, 1997.

kategorička varijabla ona se može, na način kako smo to već objasnili u odeljku **V**, kodirati kao faktor ili u kompletnu disjunktnu formu.[51]

v je kvantitativna (zavisna) varijabla
$\mathbf{x} = E \otimes v = (x_i)$ je vektor podataka, tj. niz rezultata svih ispitanika na kvantitativnoj varijabli.
Za svaki rezultat u vektoru **x** posedujemo informaciju kojoj kategoriji, tj. nivou faktora *Q* pripada. U većini statističkih programa kategorička varijabla se za analizu varijanse kodira kao **faktor**, tj. pored vektora **x** definiše se i vektor **q** u kojem je svakom ispitaniku pridružena oznaka kategorije kojoj pripada.

Klasični način definisanja i rešavanja problema :

Da li se subpopulacije definisane kategorijama varijable *Q* razlikuju u pogledu aritmetičkih sredina na kvantitativnoj (zavisnoj) varijabli v? Ako je reč o eksperimentalnom istraživanju u kojem je *Q* varijabla koja definiše nivoe eksperimentalnog faktora, onda bi formulacija ovog pitanja mogla izgledati i ovako: Postoji li efekat eksperimentalnog faktora na zavisnu varijablu?
U skladu sa teorijskim strukturnim modelom ove analize, rezultat i-tog ispitanika na varijabli v možemo posmatrati kao da je sastavljen od aritmetičke sredine svih rezultata (M), odstupanja ispitanikovog rezultata od aritmetičke sredine grupe kojoj pripada (M_k) i odstupanja aritmetičke sredine grupe od ukupne aritmetičke sredine:[52]

---

[51] Zavisno od vrste istraživanja, kategorička varijabla može biti **klasifikacioni** ili **eksperimentalni** faktor. Ako kategoričkom varijablom definišemo pripadnost ispitanika jednoj od realno egzistujućih (nezavisno od samog istraživanja) subpopulacija onda je reč o klasifikacionom faktoru. Primeri ove vrste su pol, nacionalna pripadnost, zanimanje. Ukoliko, pak, kategorička varijabla opisuje pripadnost ispitanika određenom nivou ili vrsti eksperimentalnog tretmana onda je takva varijabla eksperimetalni faktor. Primeri za to su: vrsta reči u nekom eksperimentu, doza leka u eksperimentu, vrsta metode primenjene u eksperimentu učenja.

[52] Iskazan u terminima populacionih parametara, tj. kao strukturalni model jednofaktorijalnog nacrta analize varijanse, ovaj model bi imao sledeći oblik:

$$x_{ik} = \mu + \tau_k + \varepsilon_{ik}$$

pri čemu je μ očekivana vrednost, tj. aritmetička sredina populacije u pogledu zavisne varijable, τ_k parametri koji govore o odstupanjima aritmetičkih sredina subpopulacija od aritmetičke sredine populacije ($\tau_k = \mu_k - \mu$), tj. o veličini **efekta** pojedinih nivoa faktora na zavisnu varijablu, a ε_{ij} su odstupanja individualnih vrednosti entiteta od očekivanih vrednosti subpopulacije kojoj entitet pripada ($\varepsilon_{ik} = x_{ik} - \mu_k$). Isto tako, važno ograničenje za parametre modela je:

$$\sum_{k=1}^{g} \tau_k = 0 .$$

Teorijski posmatrano, μ je nepoznata konstanta, τ_k je konstanta na k - tom nivou faktora, tj. u datoj kategoriji nezavisne varijable, a ε_{ik} je slučajna varijabla (posledica nekontrolisanih izvora variranja) koja ima normalnu funkciju gustine sa očekivanom vrednošću 0 i varijansom σ^2_ε. Tretiranje termina τ_k zavisi od toga da li je primenjen model **fiksnih** ili **slučajnih** efekata, tj. da li je faktor fiksan ili slučajan. Ako su kategorije ili nivoi faktora u istraživanju odabrani sistematski, neslučajno ili su, pak, u istraživanje uključeni svi nivoi iz skupa mogućih nivoa faktora (iz populacije nivoa za dati faktor) onda je reč o modelu **fiksnih** efekata (engl. **fixed effects model**). U slučaju kada je izbor nivoa iz populacije nivoa nekog faktora **slučajan** radi se o modelu **slučajnih** efekata (engl. **random effects model**). Tako, na primer, ako je u istraživanje uključen samo pol kao faktor sa oba moguća nivoa (muški, ženski) ili ako iz populacije padežnih oblika reči istraživač sistematski, **namerno** bira nekoliko padežnih oblika onda je reč o modelu fiksnih efekata. Kada, međutim, istraživač bira **slučajno** dva padežna oblika reči, a želi svoje zaključke da primeni na populaciju svih padežnih oblika, tada je faktor padežnog oblika **slučajan**. Ukoliko je reč o modelu sa **slučajnim** a ne **fiksnim**

$$x_{ik} = M + (M_k - M) + (x_{ik} - M_k).$$

Odavde se, budući da je $x_{ik} - M = (M_k - M) + (x_{ik} - M_k)$, ukupni varijabilitet kvantitativne (zavisne varijable) može razbiti na dve komponente:

$$\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ik} - M)^2 = \sum_{k=1}^{g} n_k (M_k - M)^2 + \sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ik} - M_k)^2$$

ili

$$SS_t = SS_b + SS_w$$

gde je:

SS_t -Suma kvadriranih odstupanja rezultata od AS svih rezultata na kvantitativnoj varijabli;

SS_b-Suma kvadriranih odstupanja aritmetičkih sredina pojedinih kategorija kategoričke varijable,tj.pojedinih grupa, od AS svih rezultata na kvantitativnoj varijabli ;

SS_w -Suma kvadriranih odstupanja rezultata od AS grupe kojoj rezultat pripada;

n_k - veličina k-tog uzorka ili grupe.

Dakle, SS_b je pokazatelj varijabiliteta na zavisnoj varijabli koji, osim od slučajnih faktora, zavisi i od toga kojoj subpopulaciji, tj. eksperimentalnom tretmanu ispitanik pripada, dok SS_w pokazuje varijabilitet koji je posledica samo slučajnih faktora.

Pretpostavke za primenu jednofaktorijalne univarijatne analize varijanse:

1. Podaci na zavisnoj varijabli v su bar sa intervalne merne skale;
2. Distribucija zavisne varijable u svakoj subpopulaciji je normalna funkcija gustine;
3. Poduzorci su izabrani nezavisnim slučajnim uzorkovanjem;
4. Homoscedascitet - varijanse subpopulacija u pogledu zavisne varijable su homogene;
5. Efekti faktora i greške na zavisnu varijablu su međusobno nezavisni ili aditivni.[53]

Postupak:

1. Na početku analize varijanse preporučljivo je testirati homogenost varijansi subpopulacija u pogledu kvantitativne varijable kao uslov za primenu jednofaktorske analize varijanse:

$$H_0: \sigma_1^2 = \sigma_2^2 = ... = \sigma_g^2$$

Najpoznatiji testovi homogenosti varijansi su Kohranov[54] (za uzorke jednake veličine), Bartletov (može se primeniti i na uzorke nejednake veličine) i Leveneov test.

---

faktorom, onda je i τ_k slučajna varijabla za koju se pretpostavlja normalna funkcija gustine sa očekivanom vrednošću 0 i varijansom σ^2_τ. Detaljnija objašnjenja ovog strukturalnog modela i njegove veze sa samim postupkom analize koji je prikazan u tekstu može se pročitati u Winer (1970), str. 56 - 62.

[53] Ova pretpostavka omogućuje da se ukupni varijablitet na zavisnoj varijabli podeli na dve aditivne, međusobno neprepokrivajuće komponente: varijabilitet koji je posledica delovanja faktora i varijabilitet greške.

Bartletov test počiva na računanju B statistika:

$$B = V / W,$$

pri čemu je

$$V = 2.303(c * \log \overline{S}_w^2 - \sum_{k=1}^{g} c_k \log S_k^2); \quad c = \sum_{k=1}^{g} c_k; \; c_k = n_k - 1; \quad \overline{S}_w^2 = \frac{\sum_{k=1}^{g} c_k S_k^2}{c}$$

dok je:

$$W = 1 + \frac{\left[\sum_{k=1}^{g} 1/c_k\right] - 1/c}{3(g-1)}$$

U prethodnim izrazima S^2_k je varijansa k-tog uzorka, a n_k je veličina k-tog uzorka. Pod pretpostavkom da su varijanse subpopulacija homogene, B statistik ima Hi-kvadrat distribuciju uzorkovanja sa g - 1 stepeni slobode. Ako je verovatnoća (pod pretpostavkom da su varijanse homogene) da dobijemo B statistik onakav kakav smo dobili ili veći od toga manja od 0.05 imamo dovoljno razloga da sumnjamo da je uslov homogenosti varijansi subpopulacija zadovoljen.
U SPSS programu je implementiran Leveneov test, a on se sastoji u jednofaktorskoj analizi varijanse rezultata ispitanika koji su iskazani kao apsolutna odstupanja rezultata od aritmetičke sredine grupe kojoj ispitanik pripada.

2. a) Ako je varijansa nehomogena (postoji heteroscedastičnost), tj. ako se nulta hipoteza

$$H_0: \sigma_1^2 = \sigma_2^2 = ... = \sigma_g^2$$

odbaci, poželjno je:

1) Odustati od analize varijanse. Ovo je posebno preporučljivo u slučajevima kada su varijanse subpopulacija heterogene, a uzorci su vrlo nejednake veličine. U tom slučaju F-test za testiranje značajnosti razlika između proseka nije nimalo **robustan**, tj. ditribucija uzorkovanja F-statistika znatno odstupa od Snidikorove distribucije;

2) obradu rezultata izvršiti nekim neparametrijskim testom (Kruskal-Wallisov test je objašnjen u glavi **XIII**) ili

3) povećati dozvoljenu verovatnoću greške α za odbacivanje tačne H_0. Na primer: novo $\alpha = 10 *$ početno α.

b) Ako su varijanse subpopulacija homogene, tj. ako je $\sigma_1^2 = \sigma_2^2 = ... = \sigma_g^2$, test statistik za testiranje nulte hipoteze

---

[54] Kohranov test podrazumeva računanje C statistika: $C = \frac{S^2_{max}}{\sum_{k=1}^{g} S_k^2}$ pri čemu je S^2_{max} najveća varijansa između g varijansi uzoraka, a S^2_k varijansa k-tog uzorka. Distribucija uzorkovanja C statistika je poznata i zavisi od broja grupa (g) i broja stepeni slobode koji je za jedan manji od veličine bilo kojeg k-tog uzorka (df = n_k-1) budući da test pretpostavlja da su uzorci jednake veličine. Ukoliko je verovatnoća da se dobije C statistik jednak ili veći od dobijenog, pod pretpostavkom da su varijanse subpopulacija homogene, veoma mala možemo pretpostaviti da su varijanse subpopulacija heterogene.

$$H_0: \mu_k = \mu \ \forall k \text{ (ili, što je isto: } \mu_1 = \mu_2 = ... = \mu_k)$$

je

$$F = S_b^2/S_w^2$$

gde su $S_b^2 = SS_b/df_b$ i $S_w^2 = SS_w / df_w$ ocene sistematske (S_b^2) i slučajne varijanse (S_w^2). Ocene varijansi S_b^2 i S_w^2 se još nazivaju i ocenama varijanse između i unutar grupa[55].
F-statistik ima pod H_0 Snidikorovu F-distribuciju uzorkovanja sa $df_b = g-1$ i $df_w = n-g$ stepeni slobode.

Dakle, suština postupka svodi se na to da se pokaže da li je ocena sistematske varijanse dovoljno veća od ocene slučajne varijanse, tako da se sistematska varijansa ne može tretirati kao posledica delovanja slučajnih faktora, već se smatra posledicom delovanja sistematskog faktora, tj. varijable *Q*.

3. Naknadni testovi: Ako F nije značajno onda se ne odbacuje H_0, što znači da se subpopulacije definisane kategorijama varijable *Q* ne razlikuju u pogledu prosečnih vrednosti na varijabli v. To znači da uzorci zapravo pripadaju istoj populaciji u pogledu merenog svojstva v.

Ako je F statistički značajan to nam govori da se subpopulacije definisane kategorijama kategoričke varijable međusobno razlikuju, ali samo na osnovu toga ne znamo koje se grupe međusobno statistički značajno razlikuju. Zato, ako je F značajno, treba utvrditi koje su razlike značajne.[56] U te svrhe mogu se upotrebiti:

a) naknadni t-testovi;[57]

b)naknadni testovi za višestruka poređenja (engl. **multiple range tests**), među kojima su najpoznatiji Duncanov, Scheffeov,Tuckeyev i Newman-Keulsov test[58].

---

[55] Mada je u jednofaktorijalnoj analizi varijanse distinkcija između modela slučajnih i fiksnih efekata čisto teorijska i ne dovodi do razlika u postupku analize, pretpostavićemo, da je reč o modelu sa fiksnim efektom faktora *Q*, jer je to u psihološkim istraživanjima uobičajeni model. U dvofaktorijalnoj (i višefaktorijalnoj) analizi varijanse razlike između ova dva modela imaju i vrlo važne praktične posledice na sam postupak analize, pre svega, na način formiranja F-količnika .

[56] Ukoliko bismo, umesto ovog postupka, uradili $k = g(g - 1)/2$, standardnih t testova, pri čemu je g broj grupa koje poredimo, verovatnoća odbacivanja nulte hipoteze, ako je ona tačna, bila bi u svakom od poređenja jednaka α, a verovatnoća da ne odbacimo tačnu nultu hipotezu bi bila $1 - \alpha$. U tom slučaju bi za svih k testova verovatnoća da ne odbacimo nultu hipotezu, ako je ona tačna, bila $(1 - \alpha)^k$. Prema tome, verovatnoća da bar za jedan par aritmetičkih sredina odbacimo H_0, čak i ako je ona u stvari tačna, bila bi $1-(1 - \alpha)^k$. To praktično znači da bi verovatnoća odbacivanja tačne nulte hipoteze primenom ovakvog, statistički pogrešnog postupka, ako je
$\alpha = 0.05$ za šest poređenja ($k = 6$), tj. pri poređenju četiri grupe ($g = 4$), bila ne 0.05 već 0.26, tj. $1 - 0.95^6$!

[57] Ovi testovi opisani su u Dragićević (1984), str. 252 - 255.

[58] Detaljan opis matematičkih postupaka za izvođenje testova višestrukih poređenja može se naći u udžbenicima koji se bave analizom varijanse (Winer, 1970; Howell, 1997).

Mere intenziteta razlika - ocene veličine efekta faktora

Korelacioni racio

Korelacioni razmer je mera povezanosti kategoričke i kvantitativne, zavisne varijable:

$$\eta^2 = SS_b/SS_t = 1 - SS_w/SS_t$$

Ovaj koeficijent je kvadrat Fišerovog eta-koeficijenta i pokazuje proporciju ukupne varijanse zavisne varijable koja je determinisana (ili je u vezi sa) subpopulacijskom pripadnošću ispitanika (npr. pripadnošću ispitanika nekom od eksperimentalnih tretmana). Ovaj razmer, drugim rečima, pokazuje proporcionalno smanjenje greške u predviđanju zavisne varijable kada znamo subpopulacijsku pripadnost ispitanika u odnosu na uspešnost u predviđanju bez poznavanja subpopulacijske pripadnosti. Fišerov eta-koeficijent (međugrupni koeficijent) je ocena intenziteta prosečnih razlika na zavisnoj varijabli onih subpopulacija koje su definisane kategoričkom varijablom.

Kvadrirana Omega

Računa se na sledeći način:

$$[SS_b - (g-1)\, S_w^2]/(SS_t + S_w^2)$$

Kvadrirana Omega je manje pristrasna ocena stvarnog efekta, tj. veličine efekta faktora u populaciji no što je to korelacioni racio.

Planirani ili aprioprni kontrasti

Ponekada istraživač, pre izvođenja samog istraživanja, ima u planu da napravi određen, manji broj poređenja između aritmetičkih sredina određenih grupa ili nivoa faktora. Takva poređenja, koja se zovu planirani ili apriorni kontrasti, mogu se napraviti čak i ako F-testom nije odbačena glavna nulta hipoteza o jednakosti aritmetičkih sredina subpopulacija. Naravno, broj takvih poređenja mora biti mnogo manji od broja mogućih poređenja.

Kontrast između g aritmetičkih sredina je posebna vrsta linearne kombinacije ili linearne funkcije tih aritmetičkih sredina, tj. linearna funkcija sa određenim ograničenjima. Opšti izraz za kontrast je: $c_1M_1 + c_2M_2 + ... + c_gM_g$, pri čemu je:

$$\sum_{k=1}^{g} c_k = 0.$$

Izborom koeficijenata $c_1, c_2, ..., c_g$ definiše se određeni kontrast. Na primer, u situaciji kada postoje tri grupe, tj. tri aritmetičke sredine, a koeficijenti su $c_1=1/2$; $c_2 = 1/2$ i $c_3 = -1$, definisani kontrast je: $(M_1 + M_2)/2 - M_3$. Dakle, u tom slučaju M_3 se poredi sa prosekom M_1 i M_2, tj. radi se o poređenju treće grupe sa prve dve grupe uzete zajedno. Ako je, pak, $c_1=1$; $c_2 = 0$, a $c_3 = -1$ onda je kontrast $M_1 - M_3$, tj. želi se poređenje aritmetičkih sredina prve i treće grupe. Testiranje nulte hipoteze za dati kontrast vrši se F-testom, a postupak računanja F-statistika koji služi tome može se pročitati u Winer, 1970.

Fišerovo izvorno definisanje i rešenje problema jednofaktorijalne analize varijanse[59]

$$S\beta = \mathbf{z}-\mathbf{g} \mid \alpha = \mathbf{g}^t\mathbf{g} = \min$$

gde je **z** varijabla v data u standardnom normalnom obliku (ako je **e** vektor jedinica, $\mathbf{z}^t\mathbf{e} = 0$, a $\mathbf{z}^t\mathbf{z} = 1$), $\mathbf{g} = \mathbf{z} - S\beta$ greška predviđanja vrednosti u **z** na osnovu $\mathbf{c} = S\beta$, a **S** selektorska matrica dobijena tako da je:

$$s_{ik} = \begin{cases} 1, & e_i \in q_k \\ 0, & e_i \notin q_k \end{cases} \quad i = 1,\cdots,n \quad k = 1,\cdots,g$$

Parcijalnom derivacijom funkcije $\alpha = \mathbf{g}^t\mathbf{g} = (\mathbf{z} - S\beta)^t (\mathbf{z} - S\beta) = 1 - 2\beta^t S^t\mathbf{z} + \beta^t S^t S\beta$ po elementima vektora β, i njenim izjednačavanjem sa nulom, dobija se

$$-2S^t\mathbf{z} + 2\, S^tS\beta = \mathbf{0}$$

tj.

$$S^tS\beta = S^t\mathbf{z}$$

Odatle je $\beta = (S^tS)^{-1} S^t\mathbf{z} = \mathbf{m} = (M_k)$ $(k = 1,\ldots,g)$, gde je **m** vektor aritmetičkih sredina grupa, $(S^tS)^{-1}$ dijagonalna matrica čiji su elementi $1/n_k$, dok je $\mathbf{c} = S\beta = S\,(S^tS)^{-1} S^t\mathbf{z} = \mathbf{Pz}$, gde je **P** projektor, a izraz **Pz** znači da u vektoru **c** kao rezultat svakog ispitanika imamo aritmetičku sredinu grupe kojoj pripada.
Budući da je $\mathbf{g} = \mathbf{z} - \mathbf{c} = (\mathbf{I} - S\,(S^tS)^{-1} S^t)\mathbf{z}$, a matrica $(\mathbf{I} - S\,(S^tS)^{-1} S^t)$ je ortogonalna na **P**, na ovaj način delimo podatke na dve ortogonalne komponente **c** i **g**.

Varijansa predviđenih vrednosti na osnovu Sβ, tj. varijansa vrednosti kriterijumske varijable projektovane u g-dimenzionalni jedinični hiperkub je $\mathbf{c}^t\mathbf{c} = \mathbf{z}^t\mathbf{Pz} = \eta^2$, a to je ujedno i kovarijansa predviđenih i empirijski dobijenih vrednosti na kriterijumskoj varijabli.
Korelacija predviđenih i dobijenih vrednosti kriterijumske varijable je tada:

$$r = \eta^2 / (\sqrt{\eta^2} * \sqrt{1}) = \eta^2 / \eta = \eta$$

Koeficijent η je Fišerov međugrupni koeficijent asocijacije.

Za testiranje hipoteze:

H_0: $\eta^2 = 0$ ili (što je isto!) H_0: $\mu_k = \mu\ \forall k$

računamo statistik:

$$H^2 = -\,[n-1-(1/2)(1+g)]\,\ln(1-\eta^2)$$

koji pod H_0 ima Hi-kvadrat funkciju uzorkovanja sa df = g-1 stepeni slobode. Izraz $(1-\eta^2)$ je zapravo Wilksova Λ.[60]

---

[59] Razumevanje Fišerovog definisanja i rešenja problema podrazumeva da su čitaocu poznati diferenciranje i optimizacija vektorskih i matričnih funkcija (osnovni elementi tog postupka dati su u Green & Carroll, 1976., str. 295-322 i u Dodatku knjizi K. Momirovića, B. Wolfa i D. A. Popovića, 1999. str. 527-541) i osnovi regresione analize. Osnovi regresione analize objašnjeni su u glavama **XI** i **XIV** ovog Priručnika. Čitalac koji ne poseduje ova znanja može ovaj deo teksta preskočiti u prvom čitanju.

[60] Wiksova Λ je detaljnije objašnjena u glavi **XVII** koja obrađuje kanoničku diskriminacionu analizu.

U ovako definisanom modelu je

$$SS_b = \sum_{k=1}^{g} n_k M_k^2 = \sum_{k=1}^{g} n_k \beta_k^2,$$

pri čemu je β_k k-ti element vektora β, a

$$SS_w = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

pa se H_0 može testirati F-testom, pri čemu je:

$$F = \frac{\frac{\eta^2}{g-1}}{\frac{1-\eta^2}{n-g}}$$

****

Jednofaktorijalna analiza varijanse se u SPSS programu može uraditi pomoću barem tri procedure:

1. **One-Way ANOVA**;
2. **Means** i
3. **Simple Factorial ANOVA** u SPSS 6.1 ili **General Linear Model** u SPSS 8.0

1. Uobičajeno se jednofaktorijalna univarijatna analiza varijanse radi u meniju **Statistics/Compare Means /One-Way ANOVA** . U dijalog prozoru **One-Way ANOVA** u okvir **Dependent list** treba uneti ime zavisne, tj. kvantitativne varijable v a u polje **Factor** ime nezavisne kategoričke varijable Q koja definiše pripadnost subpopulacijama. U polje **Define Range** treba upisati oznake kategorija (najmanje i najveće numeričke kodove koji označavaju kategorije). Pritiskom na dugme **Options** moguće je odabrati prikaz deskriptivnih statistika grupa (AS, Standardna devijacija, Standardna greška, Minimum i Maksimum) (uključena opcija **Descriptives**), testiranje homogenosti varijansi subpopulacija (uključivanjem opcije **Homogeneity of Variance**) i tražiti da se u ispisu pojavljuju nazivi kategorija (opcija **Display labels**). Pritiskom na dugme **Post-Hoc** dobija se dijalog u kome je moguće odabrati neki od mnogih naknadnih testova za multipla poređenja. Ukoliko se unapred planiraju određena poređenja onda se pritiskom na dugme **Contrasts** u polje **Coefficients** pa pritiskom na **Add** upisuju kodovi željenih planiranih kontrasta . Pri tome treba voditi računa da zbir svih koeficijenata mora biti jednak nuli i da se koeficijenti upisuju počev od kategorije označene najnižim brojem . Tako npr. ako su tri grupe označene ciframa 1, 2 i 3 a želi se poređenje grupe 1 i 3 treba upisati -1, 0, 1 a ako se želi poređenje grupe 1 naspram grupa 2 i 3 uzetih zajedno treba upisati koeficijente 1, -0.5, -0.5.

2.Budući da u ispisu iz procedure **One-Way ANOVA** nema η-koeficijenta jednofaktorijalna analiza varijanse se može uraditi i odabirom iz dijaloga **Means** (menija **Statistics/Compare Means/Means**) opcije (pritiskom na dugme **Options**) **Anova table and eta** (u okviru **Statistics for First Layer**). Inače u dijalogu **Means** pre pritiska na dugme **Options** u polje **Independent List** treba uneti ime kategoričke a u polje **Dependent List** ime kvantitativne zavisne varijable.

3. Ipak, ukoliko se želi ispis koji je u skladu sa Fišerovom postavkom i rešenjem problema potrebno je odabrati meni **Statistics/ANOVA Models/Simple Factorial ANOVA** . U dijalogu **Simple Factorial ANOVA** u polje **Dependent** unosi se ime kvantitativne varijable

a u polje **Factor(s)** ime kategoričke varijable pri čemu je neophodno popuniti opciju **Define range** u skladu sa ranije datim objašnjenjem. Pritiskom na dugme **Options** ovog dijaloga dobija se prozor u kojem treba, u okviru **Method**, odabrati opciju **Experimental** a zatim opciju **MCA** (Multiple Classification Analysis) u okviru **Statistics**.
3a. U **SPSS 8.0** jednofaktorijalna analiza varijanse može se izvesti i u proceduri **Statistics/General Linear Model**. U prozoru za dijalog **GLM – General Factorial** u polje **Dependent Variable** uneti ime zavisne varijable iz spiska u levom okviru, a u polje **Fixed Factor(s)** ime faktora, tj. nezavisne varijable. Klikom na dugme **Contrasts** ili **Post Hoc** mogu se u poddijalozima **GLM – General Factorial: Contrasts** ili **GLM – General Factorial: Post Hoc Multiple Comparisons for Observed Means** definisati željeni apriorni kontrasti ili naknadna poređenja.

Na kraju, moguće je jednofaktorsku analizu napraviti i odabirom menija **Statistics/ /Classify/Dscriminant**, tj. napraviti je kao specijalni slučaj diskriminantne analize sa jednom kvantitativnom (polje **Independents!**) i jednom kategoričkom varijablom (polje **Grouping Variable**). Ovaj postupak objašnjen je u odeljku posvećenom Kanoničkoj diskriminantnoj analizi.

Ispise iz triju odabranih procedura objasnićemo odvojeno.

1. Ispis iz procedure **One-Way ANOVA** sadrži:

Na početku ispisa ispod naslova **Oneway** date su (u koloni **Sum of Squares**) sume kvadriranih odtupanja između (red Between Groups) i unutar grupa (red Within Groups). Ocene varijansi između i unutar grupa (dobijene kao količnici suma kvadriranih odstupanja i odgovarajućeg broja stepeni slobode-kolona **DF**) nalaze se u koloni **Mean Square**. U koloni **F** je vrednost F-statistika a u koloni **Sig.** verovatnoća da F-statistik dobijen na uzorku bude jednak datoj vrednosti ili veći od nje ako je H_0 tačno.
U nastavku ispisa dati su (samo ako je u dijalogu **One-Way ANOVA:Options** uključena opcija **Descriptives**) deskriptivni statistici za svaku grupu ponaosob (veličina grupe-Count, aritmetička sredina-Means, standardna devijacija-Standard Deviation, standardna greška aritmetičke sredine-Standard Error i 95 postotni intervali poverenja za aritmetičku sredinu-95 Pct Conf Int for Mean, minimum i maximum). U koloni **Group** ovog dela ispisa će biti ili cifarske oznake grupa ili (ukoliko je u dijalogu **One-Way ANOVA Options** uključena opcija **Display Group Labels**) umesto cifarskih oznaka kategorija nazivi grupa koji su definisani kao **Value Labels** pri definisanju varijable.
Ukoliko su odabrani apriorni kontrasti (dijalogom **One-Way ANOVA:Contrasts**), posle deskriptivnih statistika grupa nalazi se matrica koeficijenata za kontraste (**Contrast Coefficient Matrix**) i rezultati t- testova kontrasta i to kako t-testa u kojem je standardna greška razlike računata na osnovu kombinovane varijanse kontrastnih grupa (**Pooled Variance Estimate**) tako i t-testa čija je standardna greška razlike izračunata na osnovu posebnih varijansi kontrastnih grupa (**Separate Variance Estimate**). Vrednost **T-prob** manja od 0.05 govori u prilog značajnim razlikama kontrastnih grupa a vrednost veća od 0.05 u prilog nepostojanju značajne razlike grupa koje su kontrastirane.
Posle Leveneovog testa homogenosti varijansi subpopulacija (**Levene Test for Homogeneity of Variances**) prikazani su pod naslovom **Multiple Range Tests** multipli testovi naknadnih poređenja (i to samo oni koji su odabrani u dijalogu **One-Way ANOVA: Post-hoc Multiple Comparisons**). U kolonama i redovima tabele koja prikazuje rezultate naknadnih testova nalaze se oznake ili nazivi svih grupa. Zvezdica u tabeli govori o postojanju značajne razlike

između grupa koje se ukrštaju u ćeliji u kojoj je zvezdica. Pod nazivom **Homogeneous Subsets** aritmetičke sredine grupa koje se ne razlikuju značajno prikazane su u okviru istog **Subset**-a.

2. U ispisu iz procedure **Means** dati su:

Pod nazivom **Descriptions of populations** deskriptivni statistici (aritmetičke sredine-kolona Mean, standardne devijacije-kolona Std dev, veličine grupa-kolona Cases i druge deskriptivne mere ako su odabrane) po grupama koji su odabrani u okviru **Cells Displays** dijaloga **Options**.
U nastavku su slično kao na početku ispisa iz procedure Oneway prikazane sume kvadriranih odtupanja između i unutar grupa, ocene varijansi između i unutar grupa i vrednost F-statistika a u koloni **Sig.** verovatnoća da F-statistik dobijen na uzorku bude jednak datoj vrednosti ili veći od n je ako je H_0 tačno. Fišerov intergrupni koeficijent dat je na kraju ispisa pod nazivom **Eta** zajedno sa kvadratom ovog koeficijenta (**Eta Squared**).

3. Ispis iz procedure **Simple Factorial ANOVA** počinje (ukoliko je u dijalogu **Simple Factorial ANOVA: Options** uključena opcija **Mean and Counts** iz okvira **Statistics**) naslovom * * * C E L L M E A N S * * * gde su date ukupna aritmetička sredina i ukupan broj ispitanika (**Total Population**) i aritmetičke sredine (i broj ispitanika) po grupama, tj. na pojedinim nivoima faktora. Budući da je ova procedura zapravo namenjena višefaktorskim nacrtima analize varijanse, sume kvadrata (**Sum of Squares**) , stepeni slobode(**DF**) i prosečni kvadrati (ocena varijanse-**Mean Square**) za glavne efekte (**Main effects**), efekte pojedinog faktora (ime kategoričke varijable koja je faktor) i za objašnjenu varijansu (**Explained**) su ovde identične i predstavljaju sumu kvadrata (stepene slobode i ocene varijanse) između grupa jer u ovom nacrtu postoji samo jedan faktor. U redu **Residual** su sume kvadrata, stepeni slobode i ocene varijanse unutar grupa. F-količnik (kolona **F**) i njegova značajnost (kolona **Sig of F**) su identični onima iz procedure **One-Way ANOVA**.
U delu ispisa M U L T I P L E C L A S S I F I C A T I O N A N A L Y S I S u kolonama **Unadjusted Dev'n** i **Adjusted for Independents Dev'n** nalaze se odstupanja aritmetičke sredina grupa (koje su definisane u redovima kolone **Variable + Category**) od ukupne aritmetičke sredine (koja je ovde data pod imenom **Grand Mean**). Ova odstupanja su u slučaju jednofaktorske analize varijanse jednaka dok u višefaktorskim nacrtima odstupanja koja su izračunata bez uzimanja u obzir drugih faktora (**Unadjusted Dev'n**) i u situaciji kada se drugi faktori statistički kontrolišu (**Adjusted for Independents Dev'n**) ne moraju biti (i obično nisu) jednaka. U koloni **eta** nalazi se Fišerov intergrupni koeficijent korelacije a u koloni **Beta** regresioni koeficijent za najtačnije (po principu najmanjih kvadrata) predviđanje vrednosti kriterijske (zavisne) varijable na osnovu pripadnosti subpopulacijama definisanim kategoričkom varijablom (nivoima faktora). U ovom slučaju su **eta** i **beta** jednaki međusobno (a i podudarni sa koeficijentom multiple korelacije -**Multiple R**) što u višefaktorskim nacrtima nije slučaj.

3a. Ispis iz procedure **GLM – General Factorial** programa SPSS 8.0 sadrži kao najosnovnije komponente ispisa:
-tabelu sa naslovom **Between-Subjects Factors** u kojoj su prikazane veličine grupa;

-tabelu **Descriptive Statistics** u kojoj su proseci i standardne devijacije na zavisnoj varijabli po pojedinim grupama;

-tabelu **Levene's Test of Equality of Error Variances** u kojoj su rezultati testiranja pretpostavke o homoscedascitetu zavisne varijable u subpopulacijama;

-tabelu **Tests of Between-Subjects Effects** u kojoj je (u redu koji nosi ime faktora) dat rezultat F-testa za testiranje nulte hipoteze o jednakosti aritmetičkih sredina subpopulacija;

-više tabela pod naslovom **Post Hoc Tests** u kojima su prikazani rezultati naknadnih testova.
Zavisno od uključenih opcija pri definisanju procedure u ispisu se mogu pojaviti i druge tabele.

Primer:

Da bi se ustanovilo postoji li efekat lišavanja snevanja na stepen uzrujanosti formirane su tri eksperimentalne grupe randomizacijom, tj.tako da su ispitanici nasumice svrstani u jednu od grupa. Podrazumeva se da su ispitanici dobrovoljno pristali da učestvuju u eksperimentu! Ispitanici u svakoj od grupa su buđeni svake noći (u određenom vremenskom periodu) podjednak broj puta a grupe su se razlikovale po stepenu u kojem su lišavane snevanja:

1. Ispitanici u grupi 1 nikada nisu bili buđeni dok sanjaju;
2. Ispitanici u grupi 2 su buđeni tri puta u trenucima kad su počinjali da snevaju a ostala buđenja su bila izvan faza snevanja.
3. Ispitanici u grupi 3 su buđeni svaki put na početku faze snevanja.

Dakle, u pogledu nivoa faktora lišavanja sna (koji je u ovom slučaju bio neponovljen jer su različiti ispitanici u svakoj od grupa) eksperimentalne grupe su bile raspoređene tako da je grupa 1 bila bez lišavanja snevanja, grupa 2 je imala mali stepen lišavanja sna a grupa 3 je bila u velikoj meri lišavana snevanja. Posle određenog broja dana ispitanicima je upitnikom izmerena uzrujanost (iritabilnost).
Na kraju ispisa iz svake od primenjenih procedura, dat je komentar ispisa.

1. Procedura One-Way ANOVA:

```
                 - - - - -  O N E W A Y  - - - - -

      Variable  UZR        SKOR NA SKALI UZRUJANOSTI
   By Variable  GRUPA      GRUPE PREMA DEPRIV. SNEVANJA

                                 Analysis of Variance

                                 Sum of          Mean            F      F
        Source            D.F.   Squares        Squares         Ratio  Prob.
Between Groups              2     985.2667       492.6333     100.9954  .0000
Within Groups              27     131.7000         4.8778
Total                      29    1116.9667

Levene Test for Homogeneity of Variances
    Statistic    df1    df2       2-tail Sig.
      1.4462      2     27          .253
```

```
- - - - - O N E W A Y - - - - -

     Variable  UZR        SKOR NA SKALI UZRUJANOSTI
  By Variable  GRUPA      GRUPE PREMA DEPRIV. SNEVANJA


 Contrast Coefficient Matrix

           Grp 1       Grp 3
                 Grp 2

Contrast  1     .0  -1.0   1.0

                                          Pooled Variance Estimate
                     Value      S. Error       T Value    D.F.       T Prob.

Contrast  1         9.5000         .9877         9.618    27.0          .000


                                        Separate Variance Estimate
                     Value      S. Error       T Value    D.F.       T Prob.

Contrast  1         9.5000        1.0609         8.954    14.3          .000

                  - - - - - O N E W A Y - - - - -

     Variable  UZR        SKOR NA SKALI UZRUJANOSTI
  By Variable  GRUPA      GRUPE PREMA DEPRIV. SNEVANJA
Multiple Range Tests:  Scheffe test with significance level .05
The difference between two means is significant if
  MEAN(J)-MEAN(I)  >= 1.5617 * RANGE * SQRT(1/N(I) + 1/N(J))
  with the following value(s) for RANGE: 3.66

   (*) Indicates significant differences which are shown in the lower triangle
                        G G G
                        r r r
                        p p p
                        1 2 3
    Mean       GRUPA
    4.4000     Grp 1
    8.6000     Grp 2     *
   18.1000     Grp 3     * *

 Homogeneous Subsets (highest and lowest means are not significantly different)
Subset 1

Group       Grp 1
Mean        4.4000
- - - - - -
Subset 2

Group       Grp 2
Mean        8.6000
- - - - - -
Subset 3

Group       Grp 3
Mean       18.1000
- - - - - -
```

Treba, pre svega, uočiti da je u ovom primeru reč o fiktivnim, tj. eksperimentalnim subpopulacijama a ne o realno egzistirajućim. Subpopulacije na koje se odnose naši zaključci su: ljudi koji normalno sanjaju, oni koji manje sanjaju no što je to prirodno i oni koji su iz nekog razloga skoro sasvim lišeni snevanja.

Na osnovu provere homogenosti varijansi subpopulacija (**Levene Test for Homogeneity of Variances**) vidimo da je uslov homogenosti ispunjen (verovatnoća da dobijemo Leveneov statistik 1.4462 ili veći od toga ako je tačna pretpostavka o homoscedastičnosti /jednakosti varijansi subpopulacija/ je 0.253).
F-statistik za testiranje nulte hipoteze o jednakosti aritmetičkih sredina eksperimentalnih subpopulacija u pogledu uzrujanosti) je 100.9954 a verovatnoća da pogrešimo odbacujući ovu nultu hipotezu ako je ona tačna je 0.0000. Stoga zaključujemo da nultu hipotezu ima smisla odbaciti i možemo reći da lišavanje snevanja ima uticaja na izraženost uzrujanosti.
Iz dela ispisa čiji je naslov **Multiple Range Tests: Scheffe test with significance level .05** vidimo da se svaka eksperimentalna grupa statistički značajno razlikuje od svake druge.
U delu ispisa sa naslovom **Contrast Coefficient Matrix** dati su koeficijenti za kontraste (ukoliko se kontrasti specifikuju pri definisanju procedure). U ovom primeru istraživač je definisao apriorni kontrast između druge i treće grupe, a t statistik za ovaj kontrast je 9.618 i statistički je značajan (**T Prob.** je .000).
Dakle, zaključujemo da postoji efekat lišavanja sna na uzrujanost. A budući da postoji značajna razlika u stepenu uzrujanosti između grupa koje su malo i koje su skoro u potpunosti lišavane sna mogli bismo zaključiti da je efekat utoliko izraženiji što je lišavanje snevanja potpunije.

---

2. Procedura Means

- - Description of Subpopulations - -

Summaries of UZR SKOR NA SKALI UZRUJANOSTI
By levels of GRUPA GRUPE PREMA DEPRIV. SNEVANJA

Variable	Value	Label	Mean	Std Dev	Cases
For Entire Population			10.3667	6.2061	30
GRUPA	1	BEZ D.S.	4.4000	1.8379	10
GRUPA	2	MALA D.S.	8.6000	2.9136	10
GRUPA	3	POTPUNA D.S.	18.1000	1.6633	10

Total Cases = 30

- - Analysis of Variance - -

Dependent Variable UZR SKOR NA SKALI UZRUJANOSTI
By levels of GRUPA GRUPE PREMA DEPRIV. SNEVANJA

Value	Label	Mean	Std Dev	Sum of Sq	Cases
1	BEZ D.S.	4.4000	1.8379	30.4000	10
2	MALA D.S.	8.6000	2.9136	76.4000	10
3	POTPUNA D.S.	18.1000	1.6633	24.9000	10
Within Groups Total		10.3667	2.2086	131.7000	30

Source	Sum of Squares	d.f.	Mean Square	F	Sig.
Between Groups	985.2667	2	492.6333	100.9954	.0000
Within Groups	131.7000	27	4.8778		

Eta = .9392 Eta Squared = .8821

Specifična informacija koju smo dobili u ovoj u odnosu na prethodnu proceduru je ocena intenziteta razlika triju subpopulacija u pogledu prosečne uzrujanosti: Eta koeficijent u ovom slučaju je 0.9392 i on govori da su razlike velike. Iz iznosa kvadrata ovog koeficijenta (0.8821) vidimo da je skoro celokupna varijabilnost u stepenu uzrujanosti determinisana eksperimentalnim faktorom tj. pripadnošću eksperimentalnim tretmanima.

---

3. Procedura Simple Factorial ANOVA- SPSS 6.1

```
                    * * *  C E L L   M E A N S  * * *
              UZR        SKOR NA SKALI UZRUJANOSTI
           by GRUPA      GRUPE PREMA DEPRIV. SNEVANJA

Total Population
    10.37
 (    30)

GRUPA

        1          2          3
     4.40       8.60      18.10
 (    10)  (    10)  (    10)

          * * *  A N A L Y S I S   O F   V A R I A N C E  * * *
               UZR       SKOR NA SKALI UZRUJANOSTI
          by   GRUPA     GRUPE PREMA DEPRIV. SNEVANJA

               EXPERIMENTAL sums of squares
               Covariates entered FIRST

                                  Sum of                  Mean              Sig
Source of Variation              Squares      DF         Square        F    of F
Main Effects                     985.267       2        492.633   100.995  .000
   GRUPA                         985.267       2        492.633   100.995  .000

Explained                        985.267       2        492.633   100.995  .000
Residual                         131.700      27          4.878
Total                           1116.967      29         38.516
30 cases were processed.
0 cases (.0 pct) were missing.

 * * *  M U L T I P L E   C L A S S I F I C A T I O N   A N A L Y S I S  * * *
              UZR       SKOR NA SKALI UZRUJANOSTI
         by   GRUPA     GRUPE PREMA DEPRIV. SNEVANJA

Grand Mean =     10.37
                                                     Adjusted for
                                       Unadjusted    Independents
Variable + Category              N     Dev'n  Eta    Dev'n   Beta
GRUPA
    1 BEZ D.S.                  10     -5.97         -5.97
    2 MALA D.S.                 10     -1.77         -1.77
    3 POTPUNA D.S.              10      7.73          7.73
                                              .94            .94
Multiple R Squared                                          .882
Multiple R                                                  .939
```

U delu ispisa iz ove procedure čiji je podnaslov **EXPERIMENTAL sums of squares Covariates entered FIRST** vidimo da je F statistik za testiranje nulte hipoteze identičan onome koji je dobijen u prethodnim procedurama jer postoji samo jedan faktor.

Odstupanja aritmetičkih sredina prve i treće grupe od zajedničke aritmetičke sredine su izraženija nego odstupanje aritmetičke sredine druge grupe (kolona **adjusted for independents Dev'n** i **Unudjusted Dev'n** u ovom su slučaju iste jer postoji samo jedan faktor). Koeficijent Beta iz Fišerovog pristupa rešavanju problema analize varijanse je 0.94 i on je u ovom slučaju (samo jedan faktor!) jednak Eta-koeficijentu. Takođe u slučaju kada je prisutan samo jedan faktor Beta je jednak koeficijentu multiple korelacije između dveju binarnih varijabli kojima bismo mogli definisati pripadnost jednoj od tri eksperimentalne grupe i uzrujanosti.

3a. Procedura GLM – General Factorial SPSS 8.0.

Univariate Analysis of Variance
Between-Subjects Factors

		Value Label	N
GRUPE PREMA DEPRIV. SNEVANJA	1	BEZ D.S.	10
	2	MALA D.S.	10
	3	POTPUNA D.S.	10

Descriptive Statistics
Dependent Variable: SKOR NA SKALI UZRUJANOSTI

GRUPE PREMA DEPRIV. SNEVANJA	Mean	Std. Deviation	N
BEZ D.S.	4.40	1.84	10
MALA D.S.	8.60	2.91	10
POTPUNA D.S.	18.10	1.66	10
Total	10.37	6.21	30

Levene's Test of Equality of Error Variances
Dependent Variable: SKOR NA SKALI UZRUJANOSTI

F	df1	df2	Sig.
1.446	2	27	.253

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a Design: Intercept+GRUPA

Tests of Between-Subjects Effects

Dependent Variable: SKOR NA SKALI UZRUJANOSTI

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Eta Squared
Corrected Model	985.267	2	492.63	100.995	.000	.882
Intercept	3224.033	1	3224.033	660.964	.000	.961
GRUPA	985.267	2	492.633	100.995	.000	.882
Error	131.700	27	4.878			
Total	4341.000	30				
Corrected Total	1116.967	29				

a R Squared = .882 (Adjusted R Squared = .873)

Post Hoc Tests

Multiple Comparisons

Dependent Variable: SKOR NA SKALI UZRUJANOSTI

Scheffe

		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
(I) GRUPE PREMA DEPRIV. SNEVANJA	(J) GRUPE PREMA DEPRIV. SNEVANJA				Lower Bound	Upper Bound
BEZ D.S.	MALA D.S.	-4.20	.988	.001	-6.76	-1.64
	POTPUNA D.S.	-13.70	.988	.000	-16.26	-11.14
MALA D.S.	BEZ D.S.	4.20	.988	.001	1.64	6.76
	POTPUNA D.S.	-9.50	.988	.000	-12.06	-6.94
POTPUNA D.S.	BEZ D.S.	13.70	.988	.000	11.14	16.26
	MALA D.S.	9.50	.988	.000	6.94	12.06

Based on observed means.

* The mean difference is significant at the .05 level.

SKOR NA SKALI UZRUJANOSTI

Scheffe

	N	Subset		
GRUPE PREMA DEPRIV. SNEVANJA		1	2	3
BEZ D.S.	10	4.40		
MALA D.S.	10		8.60	
POTPUNA D.S.	10			18.10
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed. Based on Type III Sum of Squares The error term is Mean Square(Error) = 4.878.

a Uses Harmonic Mean Sample Size = 10.000.

b Alpha = .05.

Budući da su na osnovu komentara prethodnih ispisa sasvim jasni delovi ovog ispisa daćemo samo neke napomene ovog ispisa:
-Treba uočiti u tabeli **Tests of Between-Subjects Effects** F-količnik koji se nalazi u redu GRUPA. Kao što se može i očekivati taj F jednak je F-količniku iz drugih procedura koji služi za testiranje H_0. Takođe, u istom redu u kolni **Eta Squared** nalazi se korelacioni racio, tj. ocena veličine efekta eksperimentalnog faktora GRUPA na zavisnu varijablu. U tabeli pod naslovom **Post Hoc Tests Multiple Comparisons** , u koloni **Mean Difference (I-J)** date su razlike između aritmetičkih sredina parova grupa, a u poslednjoj tabeli su rezultati Scheffe-ovog testa višestrukih poređenja. Grupe koje se ne razlikuju nalaze se u istoj koloni. U ovom primeru svaka grupa se razlikuje od svake druge i stoga ni u jednoj koloni nema više od jedne grupe.
****

X.2. Dvofaktorijalna univarijatna analiza varijanse

Eksperimentalni nacrt u kojem je svaki nivo jednog faktora, tj. svaka kategorija jedne nominalne varijable ukrštena, tj. kombinovana sa svakim nivoom drugog faktora, tj. svakom kategorijom druge nominalne varijable, a postoji samo jedna zavisna kvantitativna varijabla zahteva primenu dvofaktorijalne univarijatne analize varijanse. U ovom tekstu prikazaćemo dvofaktorijalnu analizu varijanse u kojoj su oba faktora neponovljena, tj. u kojoj je u svakoj kategoriji nominalnih varijabli (faktora) različita grupa ispitanika. Prema tome, postupak koji će biti opisan ne sme se primenjivati u situacijama kada su rezultati na različitim nivoima bilo kog faktora dobijeni od iste grupe ispitanika, tj. kada je reč o ponovljenim faktorima (repeated mesures design).[61] Isto tako, za oba faktora ćemo pretpostaviti **model fiksnih efekata.**[62]

Struktura podataka:

$E = \{e_i; i = 1,\ldots,n\} \subset P = \bigcup_{rk} P_{rk}$ je slučajni uzorak izvučen iz populacije koja predstavlja uniju s*g subpopulacija. Subpopulacije su definisane svim mogućim kombinacijama dveju nezavisnih, kategoričkih varijabli, tj. faktora;

$Q_1 = \{ q_{1r}; r = 1,2,\ldots,s\}$ je kategorička, nezavisna varijabla koja se sastoji od s kategorija, ili prvi eksperimentalni faktor sa s nivoa;

$Q_2 = \{ q_{2k}; k = 1,2,\ldots,g\}$ je kategorička, nezavisna varijabla koja se sastoji od g kategorija, ili drugi eksperimentalni faktor sa g nivoa;

v je kvantitativna (zavisna) varijabla

$\mathbf{x} = E \otimes v = (x_i)$ je vektor podataka, tj. niz rezultata svih ispitanika na kvantitativnoj varijabli. Za svaki rezultat u vektoru **x** znamo iz koje subpopulacije, tj. koje kombinacije kategorija dveju nezavisnih varijabli dolazi.
Odgovor na pitanje da li postoji efekat nezavisnih varijabli, tj. dva faktora na zavisnu varijablu dobijamo dvofaktorijalnom univarijatnom analizom varijanse.

Metoda dvofaktorijalne analize varijanse koju ćemo prikazati pretpostavlja:

1. **Normalnost** raspodele zavisne varijable za svaku kombinaciju kategorija nezavisnih varijabli, tj. u svakoj subpopulaciji koja se dobija sparivanjem svake kategorije jedne sa svakom kategorijom druge nezavisne varijable;

2. **Homogenost varijansi** zavisne varijable za svaku kombinaciju kategorija nezavisnih varijabli, tj. u svim subpopulacijam koje se dobijaju sparivanjem svake kategorije jedne sa svakom kategorijom druge varijable;

---

[61] Dobri prikazi analize varijanse za ponovljena merenja mogu se naći u Winer, 1970 i Howell, 1997.

[62] Kao što smo već objasnili u tekstu o jednofaktorijalnoj analizi varijanse (fusnota 41) fiksni faktor je onaj čije su sve moguće kategorije, tj. nivoi uključeni u istraživanje ili je, pak, od većeg broja nivoa namerno odabran samo manji broj nivoa, dok je kod slučajnih faktora u istraživanje uključen **slučajno** izabran podskup kategorija ili nivoa faktora iz mnogo većeg skupa mogućih kategorija, tj. nivoa datog faktora.

3. **Neponovljenost** oba faktora. To znači da se pretpostavlja da se (ako bi se ukrštanjem varijabli Q_1 i Q_2 dobila tabela sa s redova i g kolona) u svakoj ćeliji tabele u kojoj se ukršta red r i kolona k nalaze rezultati različite grupe ispitanika. Isto tako, ovde prikazani postupak analize podrazumeva da je u svakoj ćeliji takve tabele više od jednog ispitanika, a da su brojevi ispitanika u svim ćelijama jednaki (tzv. balansirani dizajn) ili bar da je relativni odnos broja ispitanika na različitim nivoima jednog faktora isti za sve nivoe drugog faktora. Ako faktor Q_1 ima dva nivoa, a faktor Q_2 tri nivoa, tabela sa brojevima rezultata u pojedinim ćelijama bi izgledala ovako:

	Q_2		
Q_1	n_{11}	n_{12}	n_{13}
	n_{21}	n_{22}	n_{23}

Analiza koja će ovde biti prikazana podrazumeva da je $n_{11} = n_{12} = n_{13} = n_{21} = n_{22} = n_{23}$ ili, pak, da je relativni raspored ispitanika po nivoima jednog faktora isti za sve nivoe drugog faktora, tj. da je, na primer, $n_{11} : n_{12} = n_{21} : n_{22}$, $n_{11} : n_{13} = n_{21} : n_{23}$, a $n_{12} : n_{13} = n_{22} : n_{23}$. U oba slučaja faktori Q_1 i Q_2 su nezavisni.[63] U slučaju da ovaj uslov nije zadovoljen postupak dvofaktorijalne analize varijanse koji ćemo prikazati ne bi bio adekvatan, jer tada ukupni varijabilitet na zavisnoj varijabli nije moguće razložiti na neprepokrivajuće, aditivne komponente kao u postupku koji će ovde biti prikazan. Sume kvadrata odstupanja u slučaju zavisnosti faktora zavise od redosleda kojim se faktori uvode u model.[64]

4. Za oba faktora važi **model fiksnih efekata**, tj. oba faktora su fiksna.

U jednofaktorijalnoj analizi varijanse ukupni varijabilitet (ukupna suma kvadriranih odstupanja) deli se na dva dela: unutargrupni i međugrupni varijabilitet. Deljenjem unutargrupne i međugrupne sume kvadrata odstupanja sa odgovarajućim brojem stepeni slobode dobijamo dve nezavisne ocene varijanse: S_b^2 i S_w^2. Međutim, ako su u istraživanju prisutna istovremeno dva faktora (nezavisne, kategoričke, varijable) ovakva podela ukupnog varijabiliteta na zavisnoj varijabli ne bi bila adekvatna. U slučaju kada su prisutna dva faktora, njihovim ukrštanjem dobijamo tabelu sa r redova (r = broj nivoa jednog faktora) i k kolona (k = broj nivoa drugog faktora).

---

[63] O pojmu nezavisnosti dveju kategoričkih varijabli biće reči u glavi **XII** Priručnika.

[64] O postupcima analize varijanse za različit i neproporcionalan broj ispitanika (rezultata) u pojedinim ćelijama može se pročitati u Winer (1970), str. 291 - 297.

Rezultat i-tog ispitanika u ćeliji rk može se iskazati na sledeći način[65]:

$$x_{irk} = M + (M_r - M) + (M_k - M) + (M_{rk} - M_r - M_k + M) + (x_{irk} - M_{rk}).$$

Ukupna suma kvadriranih odstupanja na zavisnoj varijabli se u tom slučaju deli na četiri komponente:

$$SS_t = SS_{br} + SS_{bk} + SS_{r \times k} + SS_{wc}$$

Pri tom je:

SS_t - ukupna suma kvadriranih odstupanja:

$$SS_t = \sum_{k=1}^{g} \sum_{r=1}^{s} \sum_{i=1}^{n_{rk}} (x_{irk} - M)^2$$

/x_{irk} je rezultat i-tog ispitanika u ćeliji rk, n_{rk} je broj ispitanika u ćeliji rk, a M je aritmetička sredina svih rezultata/;

SS_{br} - suma kvadriranih odstupanja između redova, tj. između nivoa faktora Q_1:

$$SS_{br} = \sum_{r=1}^{s} n_r (M_r - M)^2$$

/M_r je aritmetička sredina, a n_r broj ispitanika u r-tom redu, tj. u kategoriji r varijable Q_1/;

SS_{bk} - suma kvadriranih odstupanja između kolona, tj. između nivoa faktora Q_2:

---

[65] Iskazan u terminima populacionih parametara, tj. kao strukturalni model dvofaktorijalnog nacrta analize varijanse sa fiksnim efektima oba faktora, ovaj model bi imao sledeći oblik:

$$x_{irk} = \mu + \rho_r + \tau_k + \rho\tau_{rk} + \varepsilon_{irk}$$

pri čemu je μ očekivana vrednost, tj. aritmetička sredina populacije u pogledu zavisne varijable, ρ_r parametar koji govori o efektu jednog faktora ($\rho_r = \mu_r - \mu$), τ_k parametar koji govori o efektu drugog faktora ($\tau_k = \mu_k - \mu$), $\rho\tau_{rk}$ parametar za efekat interakcije dva faktora [$\rho\tau_{rk} = \mu_{rk} - (\rho_r + \tau_k + \mu) = \mu_{rk} - (\mu_r - \mu) - (\mu_k - \mu) - \mu = \mu_{rk} - \mu_r - \mu_k + \mu$], a ε_{irk} su greške tj. odstupanja individualnih vrednosti entiteta u subpopulaciji rk od očekivanih vrednosti subpopulacije kojoj entitet pripada ($\varepsilon_{irk} = x_{irk} - \mu_{rk}$). Dakle, rezultat i-tog ispitanika u ćeliji rk posmatra se kao zbir aditivnih efekata μ, ρ_r, τ_k, $\rho\tau_{rk}$ i greške ε_{irk}. Pri tom, važe sledeća ograničenja za parametre modela:

$$\sum_{r=1}^{s} \rho_r = 0;$$

$$\sum_{k=1}^{g} \tau_k = 0;$$

$$\sum_{r=1}^{s} \rho\tau_{rk} = \sum_{k=1}^{g} \rho\tau_{rk} = 0.$$

Teorijski posmatrano, μ je nepoznata konstanta, a ρ_r i τ_k su takođe, pošto se radi o fiksnim faktorima, konstante na r -tom, odnosno, k -tom nivou faktora, tj. u datoj kategoriji nezavisne varijable. Takođe $\rho\tau_{rk}$ je konstanta za datu kombinaciju nivoa dva faktora, dok je ε_{irk} slučajna varijabla (posledica nekontrolisanih izvora variranja). Za slučajnu varijablu ε_{irk} se pretpostavlja da ima normalnu funkciju gustine sa očekivanom vrednošću 0 i varijansom σ^2_ε. Model na kraju pretpostavlja nezavisnost, tj aditivnost svih efekata.

$$SS_{bk} = \sum_{k=1}^{g} n_k (M_k - M)^2$$

/M_k je aritmetička sredina, a n_k broj ispitanika u koloni k, tj. u kategoriji k varijable Q_2/;

SS_{wc} - suma kvadriranih odstupanja unutar ćelija:

$$SS_{wc} = SS_t - SS_{bc} = SS_t - (SS_{br} + SS_{bk} + SS_{r \times k})$$

/SS_{bc} - suma kvadriranih odstupanja između ćelija:

$$SS_{bc} = \sum_{k=1}^{g} \sum_{r=1}^{s} n_{rk} (M_{rk} - M)^2$$

pri čemu je M_{rk} aritmetička sredina, a n_{rk} broj ispitanika u ćeliji tabele u kojoj se sustiču ili ukrštaju r-ti red i k-ta kolona, tj. u kojoj se sustiču kategorija r varijable Q_1 i kategorija k varijable Q_2/ a

$SS_{r \times k}$ - suma kvadriranih odstupanja za interakciju između dva faktora:

$$SS_{r \times k} = SS_{bc} - SS_{br} - SS_{bk}$$

Deljenjem suma kvadrata sa odgovarajućim brojem stepeni slobode dobija se prosečni kvadrat (engl. **Mean Square**) ili ocena varijanse koja govori o efektu datog izvora varijabiliteta na varijabilitet zavisne varijable.

Pojam interakcije faktora

Postojanje interakcije među faktorima govori o tome da efekat jednog faktora na zavisnu varijablu zavisi od nivoa drugog faktora, tj. da taj efekat nije isti na svim nivoima drugog faktora. Tako, na primer, efekat vrste sporta kojom se osoba bavi na lepotu telesne konstitucije ne mora biti isti za sportiste oba pola. Moguće je da između vrste sporta i pola postoji **interakcija** u uticaju na lepotu telesne konstitucije, tj. da uticaj vrste sporta zavisi od toga da li se radi o muškarcu ili ženi sportisti.
Postojanje interakcije faktora govori o neaditivnosti glavnih efekata pojedinih faktora. Ukoliko, pak, nema interakcije među faktorima onda je njihov zajednički efekat na zavisnu varijablu aditivan, tj. jednak zbiru njihovih zasebnih efekata.

Konačna tabela dvofaktorijalne analize varijanse izgleda ovako:

IZVORI VARIRANJA	SUME KVADRATA (SS)	STEPENI SLOBODE (df)	PROSEČNI KVADRAT- OCENA VARIJANSE
FAKTOR Q_1	SS_{br}	s - 1	S^2_{br}
FAKTOR Q_2	SS_{bk}	g - 1	S^2_{bk}
INTERAKCIJA Q_1 I Q_2	$SS_{r \times k}$	(s -1) (g - 1)	$S^2_{r \times k}$
GREŠKA	SS_{wc}	n -sg	S^2_{wc}
UKUPNO	SS_t	n-1	S^2_t

Varijansa koja odgovara pojedinom izvoru varijabiliteta govori o efektu tog izvora na zavisnu varijablu. Mogući efekti su: glavni efekti svakog od faktora Q_1 i Q_2 i efekat njihove

interakcije. Odgovor na pitanje da li su u populaciji ovi efekti različiti od nule, tj. da li postoji uticaj ovih faktora na zavisnu varijablu dobijamo testiranjem statističkih hipoteza.

Testiranje statističkih hipoteza

1. Test interakcije:
Za testiranje hipoteze o nepostojanju interakcije dva faktora:

H_0: $\rho\tau_{rk} = 0$, $\forall$ rk, tj. $\mu_{rk} = \mu$, $\forall$ rk

/pri čemu je alternativna hipoteza (o postojanju interakcije):

H_0: $(\exists rk)\ \rho\tau_{rk} \neq 0$, tj. $(\exists rk)\ \mu_{rk} \neq \mu$, (Ovo se čita kao: Postoji rk za koje...ili takvo da...)/
računamo statistik:

$$F=\frac{S^2_{rxk}}{S^2_{wc}}$$

koji, ako je H_0 tačno, ima Snidikorovu funkciju gustine sa (r - 1) (k - 1) i (n - rk) stepeni slobode. Ako je verovatnoća da statistik F uzme vrednost u intervalu od 0 do one vrednosti F-a koju smo dobili manja od odabranog nivoa značajnosti odbacujemo H_0 i tvrdimo da u populaciji postoji interakcija faktora.

2. Testovi glavnih efekata faktora:

a) Faktor Q_1:
Za testiranje hipoteze da je efekat faktora Q_1 na zavisnu varijablu u populaciji jednak nuli:

H_0:$\rho_r = 0$, $\forall$ r, tj. $\mu_r = \mu$, $\forall$ r

/pri čemu je alternativna hipoteza :

H_0: $(\exists r)\ \rho_r \neq 0$, tj. $(\exists r)\ \mu_r \neq \mu$, /
računamo statistik:

$$F=\frac{S^2_{br}}{S^2_{wc}}$$

koji, ako je H_0 tačno, ima Snidikorovu funkciju gustine sa (r - 1) i (n - n_k) stepeni slobode. Ako je verovatnoća da statistik F uzme vrednost u intervalu od 0 do one vrednosti F-a koju smo dobili manja od odabranog nivoa značajnosti odbacujemo H_0 i tvrdimo da u populaciji postoji glavni efekat faktora Q_1.

b) Faktor Q_2:
Za testiranje hipoteze da je efekat faktora Q_2 na zavisnu varijablu u populaciji jednak nuli:

H_0: $\tau_k = 0$, $\forall$ k, tj. $\mu_k = \mu$, $\forall$ k

/pri čemu je alternativna hipoteza :

H_0: $(\exists k)\ \tau_k \neq 0$, tj. $(\exists k)\ \mu_k \neq \mu$, /

računamo statistik:

$$F=\frac{S^2_{bk}}{S^2_{wc}}$$

koji, ako je H_0 tačno ima Snidikorovu funkciju gustine sa (k - 1) stepeni slobode. Ako je verovatnoća da statistik F uzme vrednost u intervalu od 0 do one vrednosti F-a koju smo dobili manja od odabranog nivoa značajnosti odbacujemo H_0 i tvrdimo da u populaciji postoji glavni efekat faktora Q_2.
Po odbacivanju određene nulte hipoteze moguće je primeniti naknadna poređenja na način na koji smo to već objasnili u tekstu o jednofaktorijalnoj analizi. Isto tako, i ovde je moguće izvoditi planirane, tj. apriorne kontraste.

Zapamtite: Ukoliko odbacimo nultu hipotezu o interakciji faktora, nema uvek mnogo smisla testirati glavne efekte pojedinih faktora. Da li u situaciji kada postoji interakcija faktora ima smisla testirati efekte pojedinih faktora zavisi od prirode ove interakcije, pa odluku o tome mora u konkretnom slučaju doneti sam istraživač.

****

Budući da se dvofaktorijalna univarijatna analiza varijanse izvodi u različitim procedurama u novijim i starijim verzijama SPSS-a objasnićemo njeno izvođenje u verzijama 6.1 i 8.0 ovog programa.

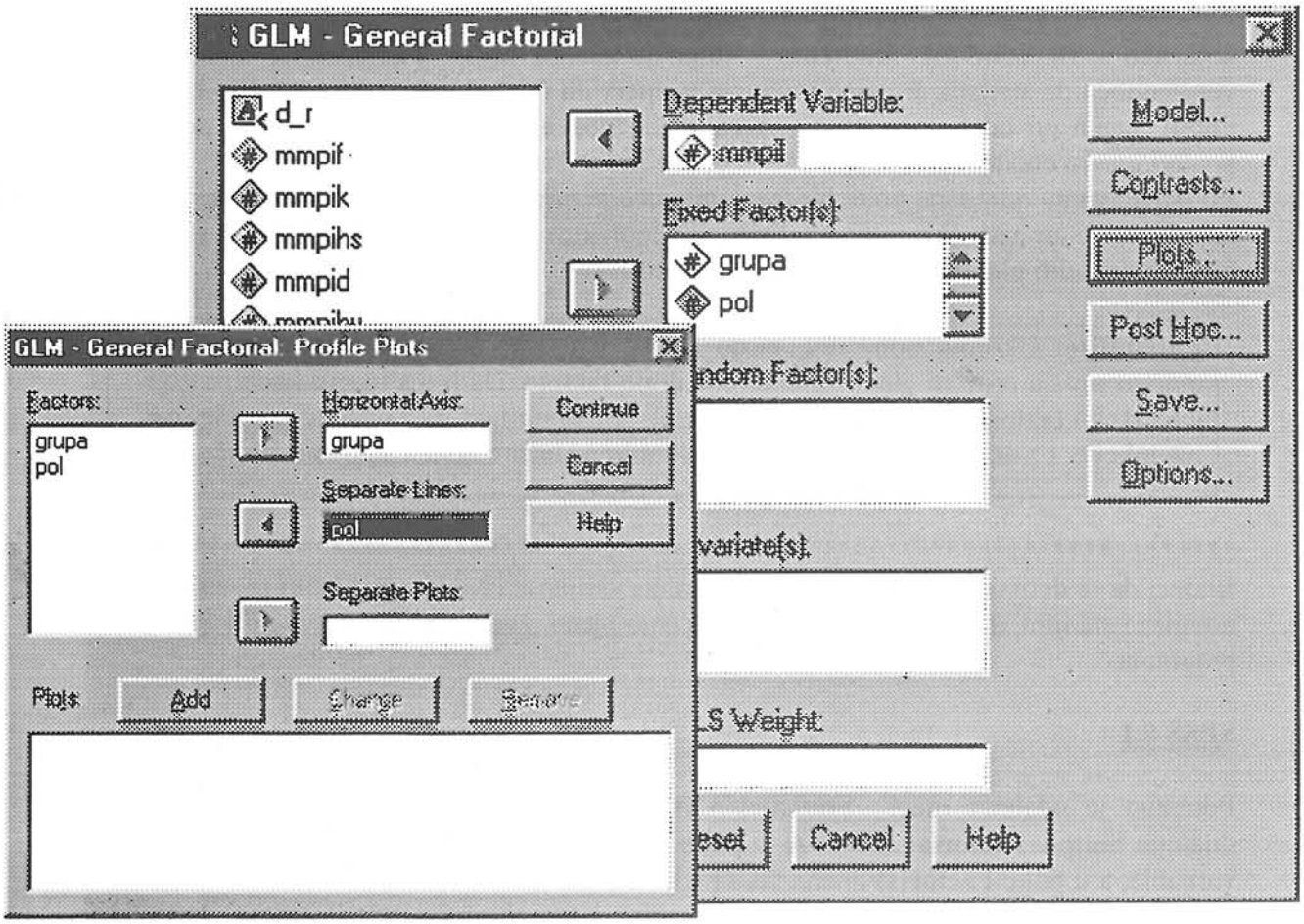
SPSS 6.1

Potrebno je odabrati meni **Statistics/ANOVA Models/Simple Factorial ANOVA**. U dijalogu **Simple Factorial ANOVA** u polje **Dependent** unosi se ime zavisne, kvantitativne varijable, a u polje **Factor(s)** imena faktora, tj. kategoričkih varijabli pri čemu je neophodno popuniti opciju **Define range** u skladu sa ranije datim objašnjenjem. Pritiskom na dugme **Options** ovog dijaloga dobija se prozor u kojem treba, u okviru **Method**, odabrati opciju **Experimental**, a zatim opciju **MCA** (Multiple Classification Analysis) u okviru **Statistics**.

SPSS 8.0

Odabrati **Statistics/General Linear Model**, u prozoru za dijalog **GLM – General Factorial** u polje **Dependent Variable** uneti ime zavisne varijable iz spiska u levom okviru, a u polje **Fixed Factor(s)** imena faktora, tj. nezavisnih varijabli (slika 7). Klikom na dugme **Plots** u poddijalogu **GLM – General Factorial: Profile Plots** moguće je definisati grafički prikaz proseka pojedinih ćelija, a klikom na dugme **Contrasts** ili **Post Hoc** mogu se u poddijalozima **GLM – General Factorial: Contrasts** ili **GLM – General Factorial: Post Hoc Multiple Comparisons for Observed Means** definisati željeni apriorni kontrasti ili naknadna poređenja. Klikom na dugme **Model** moguće je otići u poddijalog **GLM – General Factorial: Models** u kojem se može definisati željeni model analize: **Full factorial** opcija je automatski uključena i ne treba je menjati osim ako je onome ko analizira podatke sasvim jasno da želi redukovani model, tj. model u koji ne bi bili uključeni svi glavni efekti i sve interakcije faktora. Ukoliko se odabere opcija **Custom** onda treba definisati model analize kakav želimo.

Slika 7. Prozor za dijalog Generalni Linearni Model u SPSS 8.0



U primeru 2, koji ćemo objasniti u daljem tekstu (slika 7) u poddijalogu **Plots** u polje **Horizontal Axes** ubačena je varijabla Grupa, a u polje **Separate Lines** varijabla **Pol**.

Primer 1: Postoji li povezanost obolevanja od astme i pola sa jedne, i izraženosti crte depresivnosti sa druge strane? Da li se astmatičari razlikuju od neastmatičnih psihosomatskih bolesnika u pogledu prosečne depresivnosti i, ako se razlikuju, da li je ova razlika ista kod muškaraca i žena ili, pak, zavisi od polne pripadnosti?

Na slučajnom uzorku astmatičnih i neastmatičnih psihosomatskih bolesnika, muškaraca i žena, primenom upitnika Minesota multifaznog inventara ličnosti (MMPI) izmerena je izraženost depresivnosti (MMPID)kao crte ličnosti. Kao što se iz sledeće tabele vidi, klasifikacioni faktori, tj.nezavisne kategoričke varijable bile su pol (sa dve kategorije, tj. dva nivoa: muški i ženski) i kategorija psihosomatske bolesti (sa dva nivoa: pripadnost grupi astmatičnih ili neastmatičnih bolesnika). Isto tako, broj ispitanika u svakoj ćeliji tabele, tj. za svaku kombinaciji kategorija ova dva faktora je podjednak. Za oba faktora važi model fiksnih efekata.

Tabelarni prikaz uzorka po ćelijama, tj. po kombinacijama kategorija dva faktora:

GRUPA	POL: Muski	Zene	Row Total
Astm	64	64	128
Neastm	64	64	128
Column	128	128	256

Treba uočiti, isto tako, da je u ovom slučaju zapravo smisaonije depresivnost, u odnosu na astmatičnost teorijski tretirati, kao nezavisnu, a ne kao zavisnu varijablu. Ipak, bez obzira na to, odgovor na postavljeno pitanje statistički možemo dobiti dvofaktorskom univarijatnom analizom varijanse, pri čemu ćemo kao 'faktore' tretirati pol i astmatičnost, a depresivnost ćemo tretirati kao kvantitativnu (i samo pri računanju, a ne i logički) kao 'zavisnu' varijablu.

Ispis iz procedure Simple Factorial ANOVA - SPSS 6.1:

```
                * * *  C E L L   M E A N S  * * *
            MMPID
         by POL
            GRUPA

Total Population

    18.45
 (   256)

POL

        M         Z

    17.52     19.38
 (   128)  (   128)

GRUPA

      Astm.  Neastm.

    22.19     14.70
 (   128)  (   128)


          GRUPA

               Astm.  Neastm.

POL
        M      21.00     14.03
            (    64)  (    64)

        Z      23.38     15.38
            (    64)  (    64)
```

* * * ANALYSIS OF VARIANCE * * *

MMPID
by POL
GRUPA

EXPERIMENTAL sums of squares
Covariates entered FIRST

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	3806.281	2	1903.141	103.630	.00
POL	221.266	1	221.266	12.048	.001
GRUPA	3585.016	1	3585.016	195.211	.000
2-Way Interactions	17.016	1	17.016	.927	.337
POL GRUPA	17.016	1	17.016	.927	.337
Explained	3823.297	3	1274.432	69.395	.00
Residual	4627.938	252	18.365		
Total	8451.234	255	33.142		

256 cases were processed.

* * * MULTIPLE CLASSIFICATION ANALYSIS * * *

MMPID
by POL
GRUPA

Grand Mean = 18.45

Variable + Category	N	Unadjusted Dev'n	Eta	Adjusted for Independents Dev'n	Beta
POL					
Muski	128	-.93		-.93	
Zenski	128	.93		.93	
			.16		.16
GRUPA					
Astmaticari	128	3.74		3.74	
Neastmaticari	128	-3.74		-3.74	
			.65		.65
Multiple R Squared					.450
Multiple R					.671

Na početku ispisa pod naslovom **CELL MEANS** data je aritmetička sredina svihrezultata na varijabli depresivnost koja iznosi 18.45 (i ukupan broj ispitanika- 256). Zatim su dati proseci depresivnosti po kategorijama pola i astmatičnosti (GRUPA) kao i aritmetičke

sredine za svaku kombinaciju ova dva faktora. Rezultati analize varijanse dati su pod naslovom **ANALYSIS OF VARIANCE**.
Treba uočiti da je ukupna suma kvadriranih odstupanja (**Sum of Squares** u redu **TOTAL**) jednaka zbiru objašnjenog (red **Explained**) i slučajnog varijabiliteta (suma kvadrata za grešku u redu **Residual**). S druge strane, objašnjeni varijablitet se sastoji od varijabiliteta koji otpada na glavne efekte faktora (redovi **POL** i **GRUPA**) i njihovu interakciju (red **2-Way Interactions**). Isto tako, s obzirom na to da je reč o modelu fiksnih efekata F-količnici za testiranje glavnih efekata faktora i njihove interakcije dobijeni su kao količnici odgovarajućih prosečnih kvadrata za ispitivani efekat (kolona **Mean Square**) i prosečnog kvadrata za grešku (koji je u ovom slučaju bio 18.365). Iz ovog dela ispisa vidimo da postoji statistički značajan efekat pola (F=12.05) i astmatičnosti (F=195.21) ali da nema interakcije ova dva faktora u populaciji (F za interakciju pola i grupe je 0.927 i nije statistički značajan).

Deo ispisa pod naslovom **MULTIPLE CLASIFICATION ANALYSIS** smo već dobrim delom objasnili na primeru jednofaktorijalne analize varijanse. Potrebno je samo uočiti da u slučaju kada je broj ispitanika u svim ćelijama, tj. svim kombinacijama kategorija oba faktora, isti, tada nema razlike između nekorigovanih odstupanja aritmetičke sredine date kategorije faktora od ukupne aritmetičke sredine (kolona **Unadjusted Dev'n**) i odstupanja korigovanih za druge faktore (kolona **Adjusted for Independents Dev'n**). To je uvek slučaj kada su faktori nezavisni. Budući da su eta (kolona **Eta**) i beta koeficijenti (kolona **Beta**) za faktor GRUPA veći, možemo reći i da su razlike između astmatičara i neastmatičara u pogledu depresivnosti veće od razlika između muškaraca i žena, tj. da je astmatičnost bolji prediktor depresivnosti nego polna pripadnost. Izraženija depresivnost kod astmatičara no kod neastmatičara bi bila u skladu sa psihoanalitičkim tumačenjem astme kao "nesvesnog vapaja za majčinim grudima", jer bi, prema tome, depresivnost bila posledica "neprežaljenog gubitka".

Treba, takođe, uočiti da u dvofaktorijalnoj analizi varijanse multipla korelacija kategoričkih prediktora, sa jedne, i kvantitativne kriterijumske varijable sa druge strane, nije jednaka eta, niti beta koeficijentima kao što je to bio slučaj u jednofaktorijalnoj analizi. Pre svega, ovde postoje eta koeficijenti za svaki faktor i interakciju faktora, i oni su jednaki kvadratnom korenu količnika sume kvadrata odstupanja za dati efekat i totalne sume kvadrata odstupanja. Tako jeu ovom slučaju eta koeficijent za faktor GRUPA jednak 0.65, tj. jednak je kvadratnom korenu količnika sume kvadrata za taj faktor (3585.016) i ukupne sume kvadrata odstupanja (8451.234). Kvadrat multiple korelacije je (u slučaju kada je broj ispitanika u svakoj ćeliji, koja nastaje ukrštanjem svakog nivoa jednog faktora sa svakim nivoom drugog faktora, isti) jednak zbiru kvadriranih eta koeficijenta za svaki faktor i njihovu interakciju. Budući da je u ovom slučaju eta koeficijent za interakciju praktično jednak nuli, kvadrat multiple korelacije je prosto zbir kvadriranih eta koeficijenata za dva faktora.[66]
Dakle, mogli bismo na osnovu ovih rezultata zaključiti da se astmatičari razlikuju od neastmatičara u pogledu depresivnosti, kao i da je ta razlika ista u populacijama muškaraca i žena.

[66]Potpunije razumevanje ovog dela ispisa zahteva poznavanje multiple regresione analize koja je obrađena u glavi **XIV**.

Ispis iz procedure GLM - General Factorial SPSS 8.0:

Univariate Analysis of Variance

Between-Subjects Factors

		N
GRUPA	Astm.	128
	Neastm.	128
POL	Muski	128
	Zenski	128

Descriptive Statistics

Dependent Variable: MMPID

GRUPA	POL	Mean	Std. Deviation	N
Astm.	M	21,00	4,91	64
	Z	23,38	5,37	64
	Total	22,19	5,26	128
Neastm.	M	14,03	3,38	64
	Z	15,38	3,03	64
	Total	14,70	3,27	128
Total	M	17,52	5,46	128
	Z	19,37	5,91	128
	Total	18,45	5,76	256

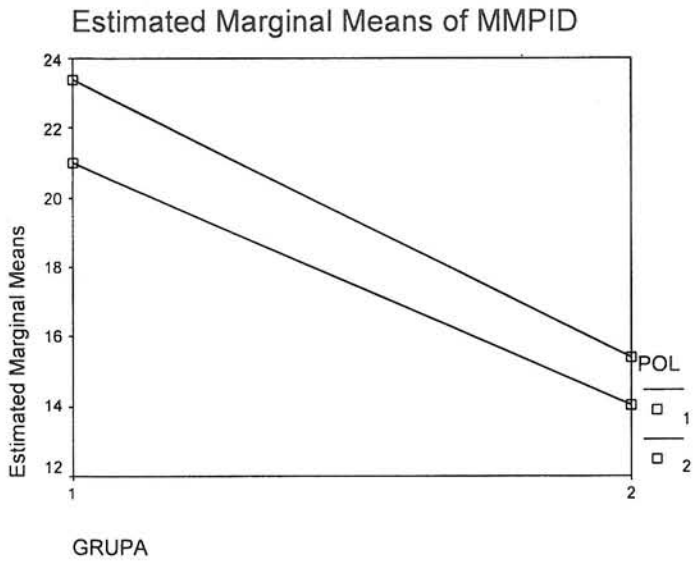
Tests of Between-Subjects Effects

Dependent Variable: MMPID

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3823,297	3	1274,432	69,395	,000
Intercept	87098,766	1	87098,766	4742,693	,000
GRUPA	3585,016	1	3585,016	195,211	,000
POL	221,266	1	221,266	12,048	,001
GRUPA * POL	17,016	1	17,016	,927	,337
Error	4627,938	252	18,365		
Total	95550,000	256			
Corrected Total	8451,234	255			

a R Squared = ,452 (Adjusted R Squared = ,446)

Profile Plots



Kako je ispis iz procedure GLM velikim delom sličan onom iz procedure Simple Factorial ANOVA dajemo samo nekoliko napomena: Sume kvadrata, varijanse i F-količnici za pojedine efekte nalaze se u redovima tabele čiji je naslov **Tests of Between-Subjects Effects**. Glavni efekti faktora su u redovima te tabele koji nose ime faktora, a greška je u redu **Error**. Grafički prikaz koji je rezultat procedure GLM (opcija **Plots**) pokazuje (**Profile Plots**) da su žene (gornja linija) prosečno depresivnije od muškaraca (donja linija). Glavni efekat pola, kao što se vidi iz tabele **Tests of Between-Subjects Effects** je statistički značajan. Takođe se vidi i na ovom grafiku da su astmatičari (grupa 1) prosečno depresivniji od neastmatičara. Budući da su nagibi linija muškaraca i žena praktično jednaki (linije su skoro paralelne), jasno je da, što smo takođe ustanovili, nema interakcije pola i astmatičnosti. Ovakav grafik je tipičan u slučajevima kada postoje glavni efekti oba faktora (linije imaju nagib i razmaknute su međusobno), a nema interakcije između faktora (linije su skoro paralelne).

Primer 2: Postoji li povezanost obolevanja od astme i pola sa jedne, i sklonosti samopredstavljanju u društveno poželjnom svetlu na upitnicima ličnosti, sa druge strane. Da li se astmatičari razlikuju od neastmatičnih psihosomatskih bolesnika u pogledu prosečne sklonosti društveno prihvatljivom odgovaranju na pitanja iz upitnika ličnosti. Ako se razlikuju, da li je ova razlika ista kod muškaraca i žena ili, pak, zavisi od polne pripadnosti?
Na slučajnom uzorku astmatičnih i neastmatičnih muškaraca i žena primenom Ajzenkovog upitnika EPQ izmerena je sklonost davanju socijalno prihvatljivih odgovora. Klasifikacioni faktori su, kao i u prethodnom primeru pol i kategorija psihosomatske bolesti (sa dva nivoa: pripadnost grupi astmatičnih ili neastmatičnih bolesnika). Isto tako, broj ispitanika u svakoj ćeliji tabele, tj. za svaku kombinaciji kategorija ova dva faktora je podjednak. Za oba faktora važi model fiksnih efekata.

Ispis iz procedure Simple Factorial ANOVA - SPSS 6.1:

* * * C E L L M E A N S * * *

EPQL
by POL
GRUPA

Total Population

11.58
(256)

POL

M	Z
11.24	11.92
(128)	(128)

GRUPA

Astm.	Neastm.
10.77	12.40
(128)	(128)

POL	GRUPA Astm.	GRUPA Neastm.
M	9.94	12.55
	(64)	(64)
Z	11.59	12.25
	(64)	(64)

* * * A N A L Y S I S O F V A R I A N C E * * *

EPQL
by POL
GRUPA

EXPERIMENTAL sums of squares
Covariates entered FIRST

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	200.195	2	100.098	11.672	.000
POL	29.566	1	29.566	3.448	.065
GRUPA	170.629	1	170.629	19.897	.000
2-Way Interactions	61.035	1	61.035	7.117	.008
POL GRUPA	61.035	1	61.035	7.117	.008
Explained	261.230	3	87.077	10.154	.000
Residual	2161.047	252	8.576		
Total	2422.277	255	9.499		

256 cases were processed.
0 cases (.0 pct) were missing.

* * * MULTIPLE CLASSIFICATION ANALYSIS * * *

EPQL
by POL
GRUPA

Grand Mean = 11.58

Variable + Category	N	Unadjusted Dev'n	Eta	Adjusted for Independents Dev'n	Beta
POL					
M	128	-.34		-.34	
Z	128	.34		.34	
			.11		.11
GRUPA					
Astm.	128	-.82		-.82	
Neastm.	128	.82		.82	
			.27		.27
Multiple R Squared					.083
Multiple R					.287

Iz tabele * * * **ANALYSIS OF VARIANCE** * * * vidimo da postoji statistički značajna interakcija pola i astmatičnosti. Značenje ove statistički značajne interakcija pola i astmatičnosti uočavamo pregledom tabele sa prosecima u pojedinim ćelijama koja se nalazi neposredno iznad naslova * **ANALYSIS OF VARIANCE** *. Iz te tabele vidimo da su muškarci "neastmatičari" prosečno skloniji 'laganju' na upitnicima od astmatičara dok sa ženama to nije slučaj. S druge strane, mogli bismo alternativno zaključiti da polne razlike u sklonosti 'laganju' postoje samo kod astmatičara, a ne i kod neastmatičnih bolesnika.

Ispis iz procedure GLM - General Factorial SPSS 8.0:

Between-Subjects Factors

		N
GRUPA	Astm.	128
	Neastm.	128
POL	M	128
	Z	128

Tests of Between-Subjects Effects
Dependent Variable: EPQL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	261.230	3	87.077	10.154	.000
Intercept	34340.723	1	34340.723	4004.477	.000
GRUPA	170.629	1	170.629	19.897	.000
POL	29.566	1	29.566	3.448	.065
GRUPA * POL	61.035	1	61.035	7.117	.008
Error	2161.047	252	8.576		
Total	36763.000	256			
Corrected Total	2422.277	255			

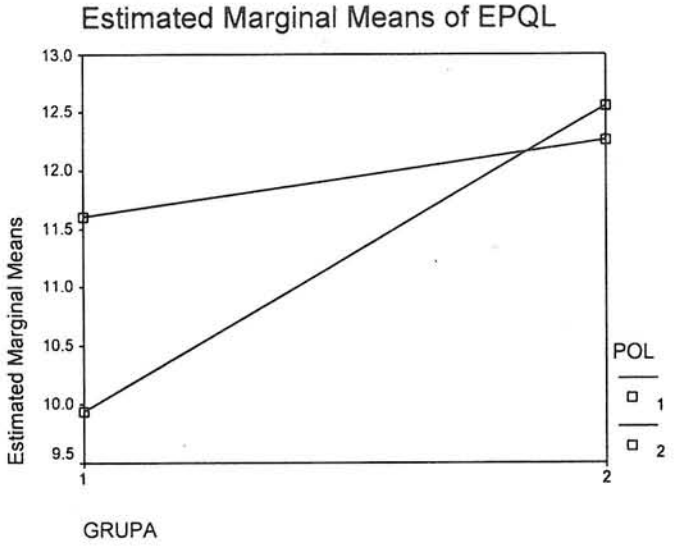
a R Squared = .108 (Adjusted R Squared = .097)

Estimated Marginal Means

GRUPA * POL
Dependent Variable: EPQL

GRUPA	POL	Mean	Std. Error	95% Confidence Interval Lower Bound	Upper Bound
Astm.	M	9.938	.366	9.217	10.658
	Z	11.594	.366	10.873	12.315
Neastm.	M	12.547	.366	11.826	13.268
	Z	12.250	.366	11.529	12.971

Profile Plots



Budući da je ispis iz procedure GLM velikim delom sličan onom iz procedure Simple Factorial ANOVA daćemo samo nekoliko napomena: Sume kvadrata, varijanse i F-količnici za pojedine efekte nalaze se u redovima tabele čiji je naslov **Tests of Between-Subjects Effects**. Glavni efekti faktora su u redovima te tabele koji nose ime faktora, a greška je u redu **Error**. Razlika ove tabele u odnosu na njoj odgovarajuću tabelu u proceduri Simple Factorial ANOVA je u iznosu Kvadrirane multiple korelacije (**R-squared**). To je posledica toga što procedura GLM SPSS 8.0 pri računanju ovog koeficijenta uzima u obzir i kvadrat eta-koeficijenta za interakciju (što je ispravno), dok procedura Simple Factorial ANOVA SPSS 6.1 uzima u obzir samo kvadrate eta-koeficijenata za glavne efekte faktora. Isto tako, grafički prikaz koji je rezultat procedure GLM (opcija **Plots**) olakšava tumačenje statistički zanačajne interakcije. Kao što se iz prethodnog grafika (**Profile Plots**) vidi, u grupi muškaraca neastmatičari su skloniji davanju socijalno poželjnih odgovora (linija sa većim nagibom), dok kod žena to nije slučaj.

XI. LINEARNA KORELACIJA I JEDNOSTRUKA REGRESIJA

Struktura podataka

$E = \{ e_i ; i = 1,....,n\} \subset P$ je slučajni uzorak izvučen iz homogene populacije

v je kvantitativna prediktorska varijabla
u je kvantitativna kriterijumska varijabla
Uzmimo da smo opisali uzorak E (tj. merenjem odredili vrednosti ispitanika) u pogledu kriterijumske varijable y, kao i u pogledu prediktorske varijable v :

$E \otimes u = \mathbf{y} = (y_i); i = 1,....,n;$
$E \otimes v = \mathbf{x} = (x_i); i = 1,....,n;$

XI.1.Linearna korelacija

Većina ljudi kada čuje termin *korelacija* zna i njegovo približno značenje. Onima koji misle da nisu čuli za račun korelacije moglo bi se reći da je to u stvari samo matematički razrađen način utvrđivanja onoga što često uviđamo u životu: neke pojave su u izvesnoj vezi (poneke čvrsto "ko Badnji dan i Božić"!, neke manje čvrsto), a neke nisu povezane. Ako između dveju varijabli postoji korelacija (povezanost) onda to znači da na osnovu vrednosti koju neko ima na jednoj varijabli možemo tačnije da predviđamo njegov reziltat na drugoj varijabli nego što to možemo bez posedavanja date informacije. Drugim rečima, korelacija između dveju varijabli pokazuje koliko informacija o jednoj varijabli je istovremeno sadržano u drugoj. Većina ljudi "zna" da ljudi koji više jedu obično bivaju deblji od onih koji su približne visine a manje jedu, da su pametnija deca generalno i uspešnija u školi... Ali ponekad se matematički dobijaju i neke neobične korelacije. Tako, istraživanjima je ustanovljeno da "postoji" korelacija između broja roda u selima i broja novorođene dece (?) ili između broja sveštenika i broja prostitutki (!).[67]

Tipovi veza između varijabli

Ako među varijablama postoji povezanost ona može biti:

-Funkcionalna: jednoj vrednosti varijable x uvek odgovara jedna određena vrednost y varijable. Na primer: y = 3x. U fizici je pritisak date mase gasa u funkcionalnoj vezi sa njenom zapreminom i temperaturom.

-Uzročno-posledična: Jedna varijabla (ili pojava) je uzrok druge. Da bi postojala kauzalna veza ona mora biti funkcionalna.

-Stohastička: jednoj vrednosti varijable x odgovara više vrednosti varijable y i to svaka sa određenom verovatnoćom. Stohastička povezanost daleko je najčešća među psihološkim varijablama.

[67] Naravno, ovde je reč o tzv. prividnim korelacijama jer je parcijalna korelacija broja roda i broja novorođene dece, ili broja sveštenika i broja prostitutki (kada se statistički kontroliše veličina mesta) jednaka nuli.

Kovarijansa i korelacija

I kovarijansa i korelacija govore o tome koliko mere jedne varijable "idu ruku pod ruku" sa merama druge varijable. Kovarijansa i korelacija pokazuju stepen kovariranja dveju varijabli, tj. stepen u kojem promene jedne varijable bivaju praćene promenama u drugoj varijabli. Drugačije rečeno, one iskazuju stepen u kojem pojedinci (ili entiteti) imaju isti relativni položaj u odnosu na ostale entitete u pogledu dveju varijabli.
Matematički posmatrano kovarijansa je prosečni krosprodukt ili unakrsni proizvod centriranih rezultata:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{n}$$

Isto tako, kovarijansa je normirani skalarni proizvod dva vektora ako su oba centrirana (od svakog rezultata oduzeta AS):

$$S_{xy} = \mathbf{d}_x{}^t \mathbf{d}_y * n^{-1}$$

gde su $\mathbf{d}_x$ i $\mathbf{d}_y$ vektori čiji su elementi centrirani rezultati na varijablama X i Y, tj. rezultati koji su iskazani kao odstupanja rezultata od njihove aritmetičke sredine.
Budući da veličina kovarijanse zavisi od metrike (tj. od merne skale) varijabli, nije jednostavno iz veličine kovarijanse ustanoviti jačinu linearne povezanosti varijabli. Zapravo, dobijenu vrednost kovarijanse treba posmatrati u odnosu prema maksimalno mogućoj kovarijansi varijabli sa datim metrikama. Najveća moguća kovarijansa među varijablama sa datim metrikama jednaka je proizvodu standardnih devijacija dveju varijabli.

Brave-Pirsonov koeficijent linearne korelacije

Koeficijent linearne korelacije matematički je jednak:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{nS_xS_y}$$

Koeficijent linearne korelacije je, prema tome, jednak količniku kovarijanse i proizvoda standardnih devijacija varijabli, tj. količniku realne i najveće moguće kovarijanse datih varijabli:

$$r_{xy} = \frac{S_{xy}}{S_xS_y}$$

Koeficijent korelacije je, očigledno, prosečni unakrsni proizvod, tj. kovarijansa

standardizovanih varijabli:

$$r_{xy} = \frac{\sum_{i=1}^{n} (z_{x_i} z_{y_i})}{n}$$

ili skalarni proizvod vektora u kojima su rezultati na varijablama x i y iskazani u standardizovanom obliku:

$$r = \mathbf{z}_x^t \mathbf{z}_y n^{-1}$$

Ako varijable predstavimo tačkama (tj. vektorima) u n-dimenzionalnom prostoru ispitanika, koeficijent linearne korelacije se može posmatrati i kao kosinus ugla između vektora varijabli ukoliko su vektori varijabli jedinične dužine.

Budući da je kosinus ugla između vektora **x** i **y** u opštem slučaju jednak:

$$\cos\alpha = \mathbf{x}^t\mathbf{y} / (|\mathbf{x}||\mathbf{y}|)$$

(gde je α ugao između vektora, dok su $|\mathbf{x}|$ i $|\mathbf{y}|$ dužine vektora **x** i **y**) jasno je da je, ako su vektori **x** i **y** jedinične dužine, kosinus ugla jednak korelaciji varijabli x i y.
Brave-Pirsonov koeficijent korelacije je indeksni broj koji govori o tipu (s obzirom na predznak) i stepenu linearne povezanosti između varijabli. Kao i kosinus ugla i koeficijent linearne korelacije se može kretati u granicama od -1 do +1. Ipak, kada se distribucije x i y varijable razlikuju po obliku, maksimalni koeficijent linearne korelacije manji je od jedinice. Koeficijenti korelacije nisu brojevi na linearnoj skali pa se ne može reći za r = 0.50 da je dva puta veće od r = 0.25. Stoga, ukoliko se želi računanje prosečne korelacije, najsmisaonije je naći kvadratni koren proseka kvadriranih koeficijenata linearne korelacije. Drugi nacin za to je da se r koeficijent pretvori u Fišerov z statistik transformacijom

$$z = 0.5 \ln [(1 + r)/(1 - r)],$$

izračuna prosečno z, a zatim dobijeno prosečno z pretvori nazad u r transformacijom:

$$r = (e^{2z} - 1) / (e^{2z} + 1),$$

gde je e = 2.71828.

Zapamtite:

- Linearnost veze među varijablama, koja je pretpostavka za računanje koeficijenta linearne korelacije, treba uvek proveriti, barem grafički (skater-dijagramom).
- Koeficijent linearne korelacije se ne menja ako se na jednu varijablu doda konstanta ili se varijabla pomnoži konstantom. Ovaj koeficijent je, prema tome, invarijantan na linearne transformacije. Dakle, korelacija između varijabli z i w, pri čemu je z = bx +c, a w = dy + e, ista je kao i korelacija između varijabli x i y. Jasno je, prema tome, da je svejedno da li ćemo korelaciju između dveju varijabli računati iz sirovih, netransformisanih rezultata ili je računamo iz devijacionih (centriranih) ili standardizovanih mera.
- Koeficijent linearne korelacije između dveju varijabli na uzorku koji je homogen u pogledu jedne od ovih varijabli manji je (zbog restrikcije raspona) nego na neselekcionisanom uzorku. Na primer, ukoliko su ispitanici selekcionisani u pogledu inteligencije tako da su na studije primljeni samo kandidati sa IQ-om iznad 120, onda će, zbog restrikcije raspona na varijabli inteligencija, koeficijent korelacije između inteligencije i uspeha na studijama biti niži no što bi bio kada bi se računao (što naravno praktično nije moguće) na uzorku svih prijavljenih kandidata.
- Koeficijent linearne korelacije ima veoma ograničenu maksimalnu vrednost (daleko manju od jedan) kada su distribucije dveju varijabli asimetrične u suprotnim smerovima ili ako je jedna od varijabli dihotomna (svedena na dve kategorije) tako da je proporcija slučajeva u kategorijama nejednaka.
- Pointbiserijski koeficijent korelacije, koji smo u Odeljku **VI.2.** opisali kao meru intenziteta razlika aritmetičkih sredina dveju subpopulacija, matematički je identičan koeficijentu linearne korelacije između kvantitativne (zavisne) varijable i binarne kategoričke varijable koja definiše pripadnost jednoj od subpopulacija .
- Zaključci o visini koeficijenata korelacije nisu tranzitivni: ako je korelacija između dveju varijabli (x i y) 0.80, a x korelira sa trećom varijablom 0.80, to nikako ne znači da je korelacija varijable y sa tom trećom varijablom 0.80. Korelacija varijable y sa trećom varijablom može biti čak 0.28.[68]
- Ukoliko na osnovu koeficijenta korelacije dobijenog na uzorku (r) **ocenjujemo** kolika je linearna korelacija između varijabli u populaciji (ρ), r je utoliko pristrasnija ocena što je uzorak manji. Nepristrasnu ocenu populacijske linearne korelacije (r_k) možemo dobiti iz r na sledeći način:

$$r_k = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{n - 2}}$$

- Tumačenje visine korelacije zavisi od konteksta u kojem se korelacija koristi. Tako, na primer korelacija od 0.80 između rezultata na nekom testu koji, po pretpostavci, meri osobinu od koje zavisi uspeh u nekoj aktivnosti i uspeha u toj aktivnosti govori o visokoj prediktivnoj valjanosti testa. Ista tolika korelacija između rezultata ispitanika u dva zadavanja istog testa (test-retest pouzdanost) biće nedozvoljeno niska!

[68] Najniža korelacija varijable y sa trećom varijablom z jednaka je $\cos 2\theta$ pri čemu je $\cos \theta = r_{xy} * r_{xz}$

Važnost poznavanja korelacije

Dva opšta tipa problema za čije je rešavanje nužno poznavanje korelacije između varijabli:

1.Praktični-Tačnost predviđanja ispitanikovog statusa na jednoj varijabli kada imamo informaciju o njegovoj meri na drugoj varijabli.

Primer: Predviđanje uspeha na poslu na osnovu testova selekcije. Ovime se bavi regresiona analiza.

2.Teorijski-Određivanje stepena u kojem se individualne razlike među ispitanicima u pogledu dveju ili više varijabli mogu pripisati istom determinišućem faktoru. Najčešće se radi o problemu opštosti pojedinih crta ili osobina ličnosti.

Primer: Postoji li generalna osobina desnorukosti?

Da bi se odgovorilo na ovo pitanje mogu se zadati različite "probe" desnorukosti:
1.Brzo i tačno zabadanje ekserčića u otvore na tabli;
2.Zakovčavanje dugmadi u datu poziciju;
3.Pakovanje blokova u kutije;
4.Zavrtanje šrafova.

Ako su korelacije među uspešnostima u obavljanju svih ovih poslova desnom rukom dovoljno visoke verovatno je da postoji jedna jedinstvena crta desnorukosti koja determiniše uspeh na svim ovim testovima. .

Tipovi povezanosti između varijabli

-Pozitivna naspram negativne povezanosti;
-Linearna naspram nelinearne, krivolinijske.

Primeri nelinearnih veza: motivacija i izvođenje neke aktivnosti, uspeh na testu inteligencije i vreme ostajanja na određenom radnom mestu.

Dijagram raspršenja (rasturanja, disperzije) - Scatter diagram

Najbolje se tip i stepen povezanosti mogu uočiti grafičkim prikazom tačaka koje su određene vrednostima entiteta (jedinica posmatranja) na dvema varijablama istovremeno, tj. koordinatama za X i Y varijablu u koordinatnom sistemu. Ovakav grafički prikaz predstavlja dijagram raspršenja. Ukoliko se pregledom dijagrama ustanovi da je generalni trend tačaka izrazito nelinearan (tj. ne formira manje ili više izduženu elipsu) onda nema smisla računanje koeficijenta linearne korelacije. Tada se za utvrđivanje (nelinearne) povezanosti među varijablama može koristiti korelacioni razmer o kojem se može pročitati u udžbenicima statistike koji obrađuju nelinearne veze među varijablama[69].

[69]Korelacioni razmer:

$$\eta_{y.x} = \sqrt{\frac{S_y^2 - S_{y.x}^2}{S_y^2}} = \sqrt{1 - \frac{S_{y.x}^2}{S_y^2}};\ \text{gde je}\ S_{y.x}^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + \ldots + n_k S_k^2}{n}$$

Testiranje statističke značajnosti koeficijenta linearne korelacije:

H_0: $\rho_{xy} = 0$ (Koeficijent linearne korelacije u populaciji jednak je 0).
H_1: $\rho_{xy} \neq 0$

Test statistik za testiranje H_0:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

ima , ako je nulta hipoteza tačna, Studentovu t-distribuciju uzorkovanja sa df = n-2.
Na osnovu toga računamo $P(|t| \geq t_{dobijeno}$ ako je H_0 tačna) = p. Ako je $p \leq 0.05$, odbacujemo nultu hipotezu i zaključujemo da u populaciji postoji linearna povezanost između varijabli. Ako je p veće od 0.05, ne odbacujemo H_0 i možemo da kažemo da nemamo razloga da tvrdimo postojanje linearne povezanosti varijabli u populaciji[70].

Zapamtite:

- Nepostojanje linearne korelacije između varijabli ili kovarijansa među varijablama jednaka nuli **ne znače** nužno i **statističku nezavisnost** dveju varijabli. Statistička nezavisnost između varijabli postoji onda kada je njihova zajednička, bivarijatna distribucija jednaka proizvodu marginalnih, tj. univarijatnih distribucija dveju varijabli. S druge strane, kada su varijable statistički nezavisne kovarijansa i linearna korelacija među njima su nužno jednake nuli.[71]
- Takođe, veoma je važno ne tumačiti korelaciju automatski kao uzročno-posledičnu vezu. Varijable, između kojih postoji povezanost, mogu zaista biti u uzročno-posledičnoj vezi, ali nam sama korelacija direktno ne kazuje da je određena veza uzročno posledična. Međutim, varijable koje su u kauzalnoj vezi moraju biti u korelacionoj Postojanje korelacije možemo tumačiti na dva načina: 1. jedna varijabla je važna komponenta druge varijable (ona čini $r^2*100\%$ komponenti druge varijable) ili 2. dve varijable imaju $r^2*100\%$ zajedničkih elemenata, tj. neki treći faktor je sadržan u obema.

U ovom obrascu S^2_y je varijansa varijable y, $S^2_{y.x}$ je prosečna (uslovna) varijansa niza vrednosti na y varijabli za pojedine vrednosti x varijable, n_1, n_2,...,n_k su brojevi rezultata na Y varijabli za svaku od k pojedinih vrednosti x varijable dok je $n = n_1+n_2+...+n_k$. Razumevanje korelacionog razmera podrazumeva poznavanje regresione analize (videti odeljak **XI.2**). Korelacioni razmer je asimetrična mera asocijacije : Korelacioni razmer u slučaju kada je y zavisna a x nezavisna varijabla nije nužno jednak korelacionom razmeru u slučaju kada je x zavisna a y nezavisna varijabla. Kao što smo već pokazali u odeljku **X**, korelacioni razmer se može koristiti i u slučaju kada je nezavisna varijabla kategorička (Fišerov intergrupni koeficijent) i nije osetljiv na permutaciju vrednosti na nezavisnoj varijabli. Korelacioni racio je jednak koeficijentu linearne korelacije u slučaju kada je regresija y varijable na x varijablu tačno linearna. Korelacioni razmer je jednak 1 ako i samo ako su x i y u striktnoj funkcionalnoj (linearnoj ili nelinearnoj) vezi. Stepen u kojem je korelacioni razmer veći od koeficijenta linearne korelacije je indikacija nelinearnosti regresije.

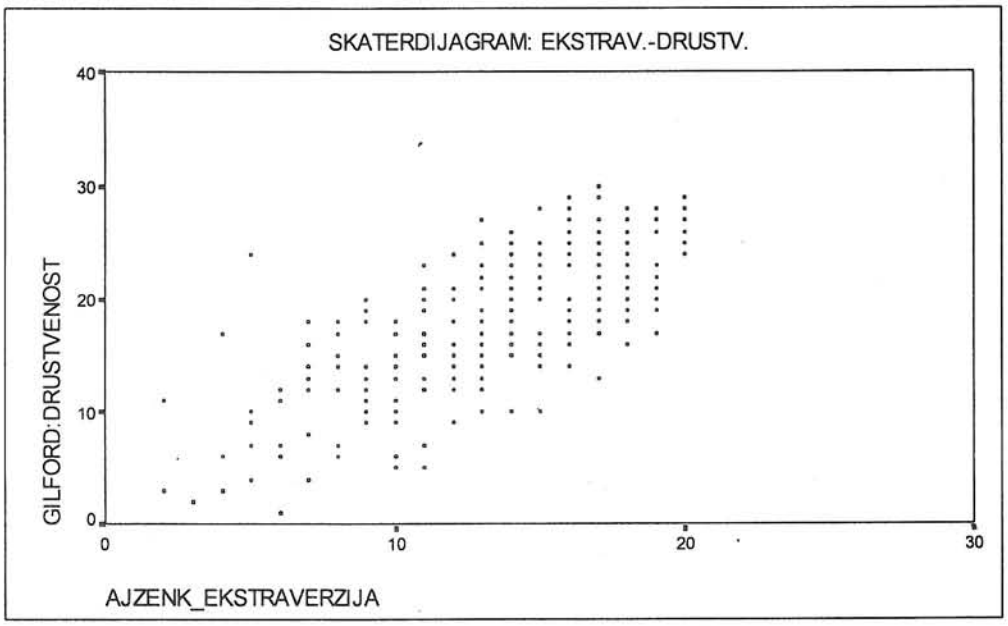
[70] Za testiranje ove hipoteze može se koristiti i statistik F koji je jednak t^2 a pod H_0 ima Snidikorovu distribuciju uzorkovanja sa $df_1=1$ i $df_2 = n-2$.

[71] Podrobnije o statističkoj nezavisnosti varijabli može se pročitati u Vuković (1997), str. 94 - 97 ili u Stilson (1966), str. 165 - 182.

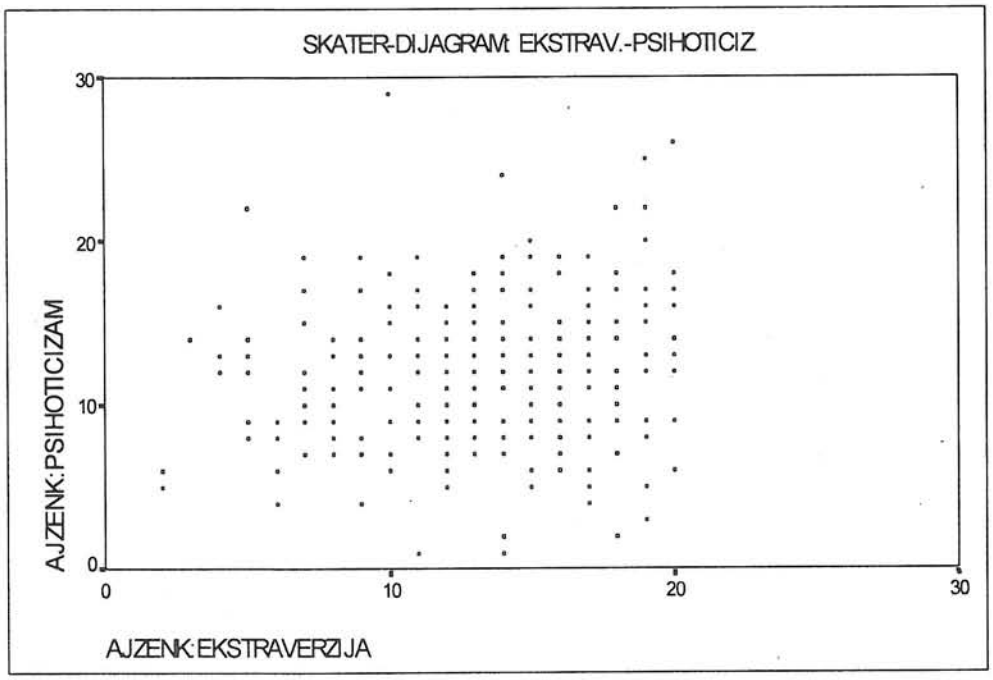
Primeri: Dobijena je korelacija između broja roda i broja novorođene dece. Prema tome, mogli bismo (pogrešno!) zaključiti da rode, ipak, donose decu! Takav zaključak bi bio fatalan po nas i po ceo ljudski rod. Naravno, uzrok ove korelacije je veličina mesta jer što je mesto veće više je i dimnjaka (roda) a i dece!
Slično je i sa korelacijom između broja sveštenika i broja prostitutki!
Medicinski primer: iz korelacije pušenja i učestalosti raka pluća ne smemo odmah zaključiti da je pušenje uzrok raka pluća (ono to može biti ali ne mora nužno, obe stvari mogu biti posledica dublje genetske ili psihološke osobine. Ipak, za neke oblike raka pluća uzročno-posledična veza sa pušenjem je prilično verovatna).

U programu SPSS se dijagram raspršenja za dve varijable može napraviti u meniju **Graphs/Scatter** tako da se u prozoru za dijalog **Scatterplot** klikne na tip **Simple** a potom na dugme **Define**. U prozoru **Simple Scatterplot** imena varijabli čija se povezanost želi da prikaže ubace se u polja **X Axes** i **Y Axes**. Ukoliko je istraživač unapred definisao jednu varijablu kao nezavisnu a drugu kao zavisnu onda se ime nezavisne varijable ubacuje u polje **X Axes** a u polje **Y Axes** ime zavisne varijable.
Po kliku na dugme **OK** u prozoru **Chart Carousel** (u verziji 6.1) ili u prozoru za ispis (u verziji 8.0) dobija se skater dijagram. Ovaj se dijagram (klikom na dugme **Edit** u prozoru **Chart Carousel**-a ili dvostrukim klikom na grafik u verziji 8.0 programa) može po želji editovati.

Grafik 11.1: Prikaz povezanosti ekstraverzije merene Ajzenkovim upitnikom EPQ i društvenosti merene Gilford-Zimmermanovim pregledom temperamenta. Između ove dve varijable postoji pozitivna linearna veza. Na grafiku se mogu uočiti i tri ispitanika koji se ne uklapaju dobro u linearni model.

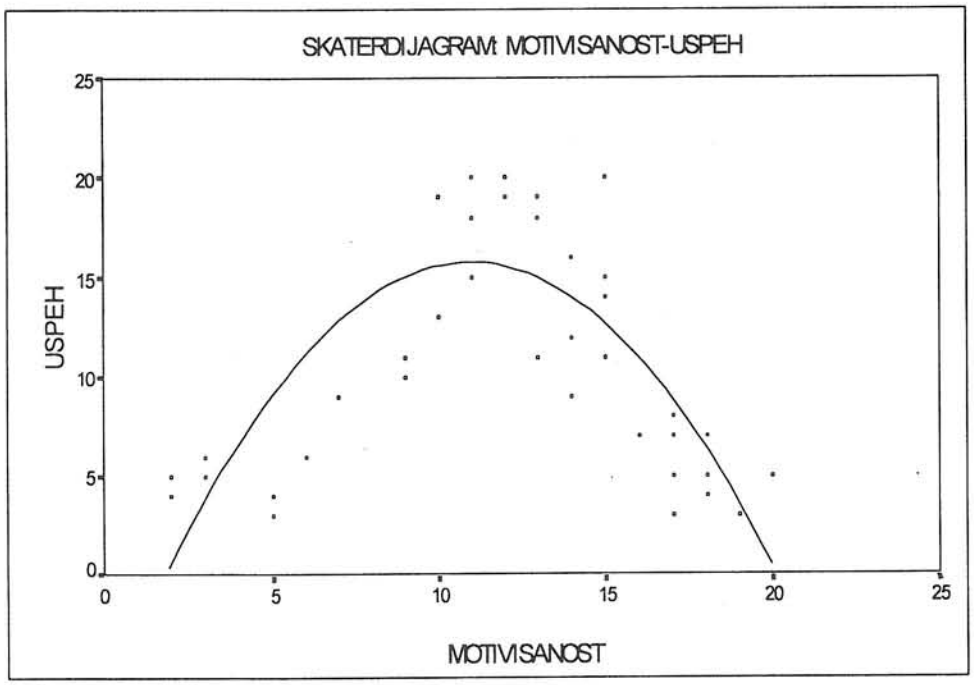


Grafik11.2: Prikaz veze između ekstraverzije i psihoticizma. Između ovih dveju varijabli očigledno nema linearne povezanosti.



Grafik 11.3: Skater-dijagram povezanosti motivisanosti i uspeha

Kriva linija na grafiku pokazuje predviđene (očekivane) vrednosti uspeha za dati nivo motivisanosti. Adekvatna mera asocijacije ovih dveju varijabli bio bi korelacioni razmer.



Računanje kovarijanse i koeficijenta linearne korelacije u programu SPSS:

Po odabiru menija **Statistics/Correlate/Bivariate**, u prozoru za dijalog **Bivariate Correlations** ubacujemo imena varijabli između kojih želimo da utvrdimo postoji li linearna povezanost. Budući da je u okviru **Correlation Coefficients** već uključena opcija **Pearson**, potrebno je (klikom na **Options**) zatražiti računanje kovarijanse (uključivanjem opcije **Cross-product deviations and covariances**). Na ispisu iz ove procedure u koloni **Variance-Covar** data je kovarijansa a u simetričnoj matrici - - **Correlation Coefficients** - - (koja je za slučaj dve varijable reda 2x2) koeficijent linearne korelacije, veličina uzorka (u zagradi) i verovatnoća greške prvog tipa, tj. verovatnoća greške koju pravimo pri odbacivanju hipoteze da je linearna korelacija u populaciji jednaka nuli ako je ova hipoteza tačna.

Na isti način (ubacivanjem većeg broja varijabli istovremeno) može se dobiti matrica interkorelacija svih varijabli. Jedina razlika, u odnosu na korelaciju između dveju varijabli, je u tome što se u opcijama (klik na dugme **Options**) može (u okviru **Missing values)** u ovom slučaju odabrati jedan od dva načina tretiranja nedostajućih podataka. Ukoliko se u tom okviru uključi opcija **Exclude cases listwise** koeficijenti korelacije između svih parova varijabli biće računati samo za ispitanike koji na svim varijablama imaju podatke. Ukoliko se, pak, uključi opcija **Exclude cases pairwise** (uključena po defaultu!), ispitanik, za koga nemamo podatke na nekoj od varijabli, biće uzet u obzir pri računanju korelacija između varijabli za koje ima podatke. Naravno, poslednja opcija može dovesti do toga da unutar iste matrice interkorelacija imamo koeficijente linearne korelacije koji su dobijeni na nejednakim uzorcima i, stoga, međusobno neuporedljivi.

Primer1: Matrica interkorelacija između ekstraverzije, psihoticizma i društvenosti.. Prvi broj u matrici je koeficijent linearne korelacije, broj u zagradi je veličina uzorka a P je, pojednostavljeno rečeno, verovatnoća da dobijemo toliki ili veći koeficijent linearne korelacije na uzorku date veličine ako u populaciji nema linearne povezanosti.

Dakle, $P = P(|t| \geq t_{dobijeno} \mid H_0 \text{ tačno})$.

Iz matrice se vidi da između društvenosti i ekstraverzije postoji visoka pozitivna linearna povezanost dok između psihoticizma i ektraverzije (i društvenosti) nema linearne povezanosti.

```
            - -  Correlation Coefficients  - -
             EXTRA        DRUS        PSIHOT

EXTRA         1.0000       .7526       .0899
             (  282)     (  282)     (  282)
             P= .        P= .000     P= .132

DRUS           .7526      1.0000       .0335
             (  282)     (  282)     (  282)
             P= .000     P= .        P= .575

PSIHOT         .0899       .0335      1.0000
             (  282)     (  282)     (  282)
             P= .132     P= .575     P= .
(Coefficient / (Cases) / 2-tailed Significance)
" . " is printed if a coefficient cannot be computed
```

Primer2: Između motivisanosti i uspeha u obavljanju date aktivnosti nema **linearne** povezanosti. Ispis iz procedure **Corelate / Bivariate** za ove dve varijable pokazuje da je koeficijent linearne korelacije 0.06 i da nije statistički značajan. Međutim, ako relaciju ovih dveju varijabli prikažemo skaterdijagramom (videti Grafik3), uočavamo da **postoji**

povezanost ovih dveju varijabli ali je ta veza (u skladu sa Jerks-Dodsonovim zakonom) **nelinearna**[72]. Ukoliko varijablu motivisanosti iskažemo kao
(a + motivisanost + motivisanost**2, gde je a konstanta iz regresionog modela /videti o regresiji u daljem tekstu/, *motivisanost* linearna, a *motivisanost**2* kvadratna komponenta) dobijamo koeficijent korelacije između tako definisane motivisanosti i uspeha od 0.79!

```
- -  Correlation Coefficients  - -

                MOTIV          USPEH

MOTIV          1.0000          .0624
               (   40)        (   40)
               P= .           P= .702

USPEH           .0624         1.0000
               (   40)        (   40)
               P= .702        P= .
```

XI.2.Jednostruka linearna regresija

Termin regresija ima u psihologiji različita značenja.[73] U ovom tekstu, termin (jednostruka linearna) regresija biće korišćen u statističkom smislu, tj. u smislu pronalaženja najadekvatnijeg **linearnog** modela za predviđanje vrednosti kriterijumske (zavisne) varijable na osnovu poznavanja rezultata ispitanika na prediktorskoj (nezavisnoj) varijabli. Cilj regresione analize kao statističke tehnike je, dakle, nalaženje jednačine koja dobro opisuje vezu između kriterijumske i prediktorske varijable da bi se na taj način, na osnovu vrednosti ispitanika na prediktorskoj varijabli mogao što tačnije predvideti njegov rezultat

---

[72] Yerks-Dodsonov zakon nastao je na osnovu istraživanja na pacovima u kojima je nađena inverzna U veza između jačine nagona (motivacije) i uspešnosti izvođenja zadatka.Osim toga, prema ovom "zakonu" optimalni nivo motivacije za dati zadatak opada sa težinom zadatka. Što je, dakle, zadatak složeniji to niži stepen motivacije daje najbolje izvođenje. Za proste zadatke, jača motivacija znači i bolje izvođenje. Za složenije zadatke to važi samo do određene tačke. Povećanje motivacije dalje od te tačke dovodi do smanjenja uspešnosti izvođenja. Prema H. Ajzenku sličan odnos postoji i između nivoa pobuđenosti (arousal) i uspešnosti izvođenja zadatka.

[73] Doslovno značenje ove reči je kretanje unazad (lat. regressio). Ona u psihologiji označava i vraćanje na raniji ili manje zreo način ponašanja ili nivo organizacije ponašanja. U psihoanalitičkoj teoriji to je mehanizam odbrane koji se sastoji u tome da se osoba (suočena sa instiktivnom željom ili impulsom koji dovode do anksioznosti) vraća na stadijum koji je ranije bio katektiran libidom i tako manje provocira anksioznost i osećanje krivice. Dva oblika ovog vraćanja postoje: a) vraćanje na raniji objekat; b) vraćanje na nivo infantilne seksualne organizacije.
Statistički, regresija označava takvu vezu između dve varijable u kojoj je predviđeni skor zavisne varijable bliži AS zavisne varijable (regredira ka njoj) nego što je njoj odgovarajuća vrednost nezavisne varijable blizu AS nezavisne varijable. Ime ovoj statističkoj metodi poteklo je na osnovu zapažanja Darvinovog učenika Sir Fransisa Goltona do kojih je došao ispitujući vezu između visine očeva i sinova. Golton je primetio da, osim što, prosečno gledano, visoki očevi imaju visoke sinove (što bi se i očekivalo) postoji i jedna neobična stvar: ekstremno visoki (ili niski) očevi imaju sinove koji su (u odnosu na prosek) manje ekstremno visoki (ili niski). Na osnovi ovih zapažanja Golton je pretpostavio postojanje prirodnog "zakona regresije", tj.regresije (vraćanja) ka prosečnosti. Potom je među statističarima Goltonov termin regresija zauzeo mesto matematičkog termina funkcija pa je funkcija y = ax + b postala regresiona jednačina.

na kriterijumskoj varijabli onda kada nismo u mogućnosti za nekog ispitanika da rezultat na kriterijumskoj varijabli dobijemo merenjem. Praktično, radi se o tome da se na osnovu podataka na prediktorskoj i kriterijumskoj varijabli dobijenih na reprezentativnom uzorku iz neke populacije specifikuje jednačina čiji je opšti oblik y = f(x) /čitamo kao: y je funkcija od x/ koja će poslužiti u prediktivne svrhe. Drugim rečima, kroz raspored tačaka (definisanih koordinatama empirijskih vrednosti kriterijumske i prediktorske varijable) u dijagramu raspršenja treba provući pravu liniju koja, najčešće po principu najmanjih kvadrata, najbolje aproksimira dati raspored tačaka. Prema kriterijumu najmanjih kvadrata (least squares) zbir kvadriranih udaljenja tačaka od tako provučene (regresione) prave treba da bude manji no od bilo koje druge prave. Da bismo odredili gde treba provući pravu liniju, budući da je jednačina prave linije y = a + bx, potrebno je odrediti dva parametra (tj. oceniti ih na osnovu empirijskih podataka):
a) intercept , tj. a iz gornje jednačine i
b) nagib, tj. b iz gornje jednačine[74].

Kada regresionu pravu definišemo na osnovu položaja empirijski dobijenih tačaka (svaka tačka predstavlja jedinicu posmatranja- najčešće je to ispitanik a pozicija tačke određena je rezultatima ispitanika na varijablama x i y), jednačina prave linije dobija oblik:

$$y^*_i = a + bx_i$$

gde je y^*_i predviđeni ili očekivani rezultat i-tog ispitanika na kriterijumskoj varijabli na osnovu njegovog rezultata na x, tj. prediktorskoj varijabli.

Formalna postavka i rešenje problema jednostruke linearne regresije

Kriterijumska i prediktorka varijabla nestandardizovane:

$$\mathbf{y} = a + b\,\mathbf{x} + \mathbf{g} \mid \mathbf{g}^t\mathbf{g}n^{-1} = \sum g_i^2\, n^{-1} = \text{minimum}$$

Izraz g_i^2 je kvadrirana razlika predviđene (y_i^*) i empirijski dobijene vrednosti (y_i) kriterijumske varijable, kvadrirana greška predviđanja ili kvadrirani rezidual.

Rešenje po principu najmanjih kvadrata ($\sum g_i^2\, n^{-1}$ = minimum):

$$a = M_y - bM_x$$

(M_y i M_x su aritmetičke sredine x-prediktorske i y-kriterijumske varijable).

$$b = S_{xy}/S_x^2 = r^*(S_y/S_x) = \sum d_{yi}d_{xi}/(\sum d_{xi}^2)$$

(S_{xy} je kovarijansa prediktorske i kriterijumske varijable, a S_x^2 varijansa prediktorske varijable, S_y je standardna devijacija kriterijumske varijable, S_x je standardna devijacija

[74] Jednačina prave linije u opštem slučaju je y=a + bx, pri čemu su a i b konstante, tj. parametri. Konstanta a je intercept, tj odsečak na Y osi za x = 0, dok je b koeficijent pravca ili nagib linije u odnosu na x osu. Ovaj koeficijent jednak je tangensu ugla koji prava linija zaklapa sa X osom

prediktorske varijable a d_{yi} i d_{xi} su devijacioni (centrirani) skorovi za i-tog ispitanika. Koeficijent a je očekivana, tj.predviđena vrednost kriterijumske varijable kada je vrednost prediktorske varijable jednaka 0. Budući da često vrednost 0 na prediktorskoj varijabli nema smisla to ni koeficijent a nema uvek smislenu sadržinsku interpretaciju.
Koeficijent b je nagib regresione linije, tj. prosečna promena kriterijumske varijable za promenu prediktorske varijable za jednu jedinicu. Ovaj koeficijent je, dakle, mera predviđene brzine promene kriterijumske varijable.

Predviđene vrednosti kriterijumske varijable y_i^* se dobijaju kao:

$$y^*_i = a + bx_i = M_y + r\,(S_y/S_x)\,(x_i - M_x)$$

Kriterijumska i prediktorska varijabla standardizovane:

$$\mathbf{z}_x b^* = \mathbf{z}_y + \mathbf{g} \mid \sigma^2_g = \mathbf{g}^t\mathbf{g} * 1/n = \text{minimum}$$

($\mathbf{g}$ je vektor grešaka /$\mathbf{g} = \mathbf{z}_y - \mathbf{z}_x\, b^*$/, tj. reziduala a b^* je standardizovani regresioni koeficijent, dok je a koeficijent u ovom slučaju jednak 0).

Rešenje:

$$b^* = r_{xy}$$

Standardizovani regresioni koeficijent pokazuje za koliko se standardnih devijacija prosečno menja y kada se x varijabla promeni za jednu standardnu devijaciju.
Predviđene vrednosti kriterijumske varijable (z_{y*i}) na osnovu standardizovanih vrednosti prediktorske varijable dobijaju se po sledećem obrascu:

$$z_{y*i} = r * z_{xi}$$

Iz prethodnog obrasca može se uočiti, budući da je r manje ili jednako 1, da je predviđena vrednost kriterijumske varijable najčešće bliža aritmetičkoj sredini (tj. 0) nego što je vrednost kriterijumske varijable u odnosu na aritmetičku sredinu stndardizovane prediktorske varijable (tj.0). Isto tako, budući da je predviđena kriterijumska varijabla dobijena množenjem vrednosti na standardizovanoj prediktorskoj varijabli koeficijentom linearne korelacije (tj. konstantom), aritmetička sredina predviđene kriterijumske varijable je 0 (tj. jednaka aritmetičkoj sredini standardizovane prediktorske varijable), a varijansa ove varijable je r^2 (budući da je r^2 pomnožena sa varijansom standardizovane prediktorske varijable, tj sa jedinicom)[75]. Kako je $z_{yi} = z_{y*i} + g_i$, to je i aritmetička sredina grešaka predviđanja, tj. reziduala jednaka 0.

Odnos standardizovanog i nestandardizovanog regresionog koeficijenta:

$$b^* = b\,(S_x/S_y)$$
$$b = b^*\,(S_y/S_x) = r_{xy}\,(S_y/S_x)$$

[75] Ovo sledi na osnovu poznavanja uticaja množenja konstantom na aritmetičku sredinu i varijansu (o čemu je bilo reči u odeljcima **IV.4.** i **IV.5**).

Određivanje podešenosti modela podacima (goodness of fit) i uspešnosti u predviđanju kriterijumske varijable

Kada se iz empirijskih podataka dobije, po principu najmanjih kvadrata, jednačina prave linije ili regresiona jednačina, potrebno je utvrditi njenu prilagođenost podacima ili njenu podesnost kao modela podataka, tj. preciznost u predviđanju vrednosti kriterijumske varijable na osnovu vrednosti prediktorske varijable.
Jedan od pokazatelja podesnosti dobijenog linearnog modela je proporcija ukupne varijanse kriterijumske varijable koja je objašnjena (ili se može predvideti) na osnovu varijacija na prediktorskoj varijabli.
Naime, ukupni varijabilitet kriterijumske varijable (definisan sumom kvadriranih odstupanja pojedinačnih rezultata od aritmetičke sredine na kriterijumskoj varijabli y i označen sa SS_t) možemo razbiti na dve komponente:

$$\sum_{i=1}^{n}(y_i - M_y)^2 = \sum_{i=1}^{n}(y_i^* - M_y)^2 + \sum_{i=1}^{n}(y_i^* - y_i)^2$$

ili

$$SS_t \quad = \quad SS_r \quad + \quad SS_e$$

gde je:

SS_t - Suma kvadriranih odstupanja rezultata od AS svih rezultata na kriterijumskoj varijabli tj. ukupna suma kvadriranih odstupanja;

SS_r - Suma kvadriranih odstupanja za regresiju, tj. suma kvadriranih odstupanja predviđenih (na osnovu regresionog modela) vrednosti kriterijumske varijable od AS rezultata na kriterijumskoj varijabli. Naravno, aritmetička sredina dobijenih i predviđenih vrednosti kriterijumske varijable je ista.

SS_e - Suma kvadriranih odstupanja predviđenih od empirijski dobijenih rezultata na kriterijumskoj varijabli, suma kvadrata grešaka predviđanja (ili reziduala).

Proporcija ukupnog varijabiliteta kriterijumske varijable koja se može predvideti (objasniti) varijacijama na prediktorskoj varijabli je koeficijent determinacije:

$$r_{y.x}^2 = \frac{SS_r}{SS_t} = \frac{\sum_{i=1}^{n}(y_i^* - M_y)^2}{\sum_{i=1}^{n}(y_i - M_y)^2}$$

Istovremeno, kada su kriterijumska i prediktorska varijabla standardizovane proporcija objasnjene varijanse kriterijumske varijable jednaka je varijansi predviđene kriterijumske varijable jer je tada proporcija objašnjene varijanse: $r^2/1 = r^2$.
Što je proporcija ukupnog varijabliteta koja je objašnjena varijabilitetom prediktora veća (bliža 1) to je model bolje prilagođen podacima[76].

[76] Dakle, kao valjanu meru podesnosti modela treba koristiti r^2, a ne r, jer je $r^2+1 - r^2 = 1$, dok , očigledno, $r+\sqrt{(1-r^2)} \neq 1$.

Zapamtite:

- Korelacija predviđene i empirijski dobijene kriterijumske varijable jednaka je koeficijentu linearne korelacije empirijski dobijene kriterijumske varijable i prediktorske varijable.[77]
- Varijansa predviđene kriterijumske varijable uvek je **manja** ili **jednaka** varijansi empirijski dobijene kriterijumske varijable[78].
- Kada su kriterijumska i prediktorska varijabla standardizovane varijansa reziduala (S_g^2) jednaka je 1- r^2 a varijansa reziduala je tada istovremeno i proporcija varijanse kriterijumske varijable koja nije objasnjena varijacijama u prediktorskoj varijabli[79].
- Korelacija prediktorske varijable i reziduala jednaka je nuli[80].

Mere tačnosti predviđanja vrednosti kriterijumske varijable na osnovu prediktora

Standardna greška predviđanja (ili ocene)

Standardna greška je mera podesnosti modela koja je zasnovana na rezidualima, tj. varijansi reziduala:

$$\sigma_g^2 = \sigma_y^2 = \rho^2\sigma_y^2 = \sigma_y^2(1-\rho^2)$$

Standardna greška predviđanja (σ_g) je standardna devijacija distribucije reziduala (grešaka predviđanja), tj. mera prilagođenosti regresionog modela podacima i istovremeno mera prosečne(verovatne) greške u predviđanju. Ocena standardne greške na osnovu podataka na uzorku dobija se na sledeći način:

$$SE_{y.x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-2}(n-1)} = S_y\sqrt{(1-r_{y.x}^2)\frac{n-1}{n-2}}$$

Budući da je standardna greška ocene statistik koji dobijamo na osnovu ocene dva parametra, broj stepeni slobode je za dva manji od broja parova rezultata na dvema

---

[77] Ovo sledi iz činjenice da se r ne menja kada se jedna varijabla linearno transformiše. a predviđena kriterijumska varijabla (y^*) je samo linearna transformacija ($y^* = a + bx$) prediktorske varijable.

[78] Kako je $r^2 \leq 1$ a $S_{y^*}^2 = r^2$, to je $S_{y^*}^2 \leq S_y^2$.

[79] Budući da je aritmetička sredina reziduala u tom slučaju jednaka nuli,

$$S_g^2 = \frac{\sum_{i=1}^{n}(g_{yi}-0)^2}{n} = \frac{\sum_{i=1}^{n}(z_{yi}-z_{y^*i})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(z_{yi}-r*z_{xi})^2 = 1-2r^2+r^2 = 1-r^2$$

[80] Pokazaćemo ovo u slučaju kada su kriterijumska, prediktorska varijabla i reziduali standardizovani. Tada je:

$$r_{xg} = \frac{\sum_{i=1}^{n} z_x z_g}{n} = \frac{\sum_{i=1}^{n} z_x(z_y - r*z_x)}{n} = \frac{1}{n}\sum_{i=1}^{n}(z_x z_y - rz_x^2) = r - r = 0$$

varijablam (n-2). Za velike uzorke, izraz (n-1)/(n-2) se bliži jedinici pa se na velikim uzorcima ovaj izraz i ne uzima u obzir pri računanju standardne greške ocene. Iz ovog obrasca može se videti kolika treba da bude korelacija prediktora i kriterijuma da bi se standardna greška ocene smanjila za određeni procenat. Na primer, da bismo standardnu grešku predviđanja prepolovili, koeficijent linearne korelacije između x i y varijable mora biti 0.866!

Standardna greška može da varira između 0 i standardne devijacije kriterijumske varijable. Iz obrasca za njenu računanje je jasno da je standardna greška ocene jednaka standardnoj devijaciji kriterijumske varijable ako je korelacija između prediktorske i kriterijumske varijable jednaka 0 a da je jednaka 0 kada je korelacija između prediktorske i kriterijumske varijable 1. Budući da je izražena u mernim jedinicama kriterijumske varijable standardna greška nije metrički invarijantna što, u slučaju kada je kriterijumska varijabla nestandardizovana, otežava poređenje standardnih greški za kriterijumske varijable sa različitim metrikama. U tom slučaju, bolje je koristiti relativne mere odstupanja empirijskih podataka od regresione linije kao što je odnos objašnjenog i ukupnog varijabiliteta, tj. koeficijent determinacije.

Testiranje hipoteza o regresionim parametrima:

Statistici a i b su ocene parametara α i β. Ovde će biti prikazano samo testiranje hipoteza da su parametri α i β jednaki 0.

Testiranje hipoteze o parametru α:

Za testiranje nulte hipoteze:

$$H_0: \alpha = 0$$

računamo statistik:

$$t = (a - \alpha) / SE_{\alpha},,$$

pri čemu je:

$$SE_a = \sqrt{\frac{SE_{y.x}}{n} - \frac{M_x^2 SE_{y.x}^2}{\sum_{i=1}^{n}(X_i - M_x)^2}}$$

Statistik t, pod H_0 (budući da je $\alpha = 0$), jednak je:

$$t = a / SE_a$$

i ima Studentovu t-distribuciju uzorkovanja sa df = n - 2.

Testiranje hipoteze o parametru β

Standardna greška b-koeficijenta je:

$$SE_b = \frac{SE_{y.x}}{S_x\sqrt{n-1}}$$

Za testiranje nulte hipoteze:

$$H_0: \beta = 0$$

računamo statistik:

$$t = (b - \beta)/ SE_b$$

koji je pod H_0 (budući da je $\beta = 0$) jednak:

$$t = b / SE_b$$

i ima Studentovu t-distribuciju uzorkovanja sa df = n-2.
Test značajnosti b-koeficijenta u jednostrukoj regresiji identičan je testiranju značajnosti koeficijenta linearne korelacije između prediktorske i kriterijumske varijable.

Veza između korelacije i regresije

Ako su prediktorska i kriterijumska varijabla standardizovane koeficijent linearne korelacije između njih jednak je kosinusu ugla koji zaklapaju regresione linija za predviđanje x na osnovu y i za predviđanje y na osnovu x. Ove dve regresione linije seku se u tački čije su koordinate jednake aritmetičkim sredinama dveju varijabli, tj. u tački čije su standardizovane koordinate (0;0).
Koeficijent linearne korelacije jednak je geometrijskoj sredini koeficijenata b (za predviđanje x na osnovu y i za predviđanje y na osnovu x):

$$r = \sqrt{(b_{yx} * b_{xy})}$$

Ukoliko sve tačke u dijagramu raspršenja padaju tačno na pravu liniju ove dve regresione linije će se poklopiti, predikcija jedne varijable na osnovu druge biće savršena, regresioni koeficijenti b_{yx} i b_{xy} će biti jednaki, a korelacija će biti 1 ili -1. Što je korelacija između varijabli niža to će ugao pod kojim se seku dve regresione linije biti veći a kada su varijable nezavisne (r = 0) ugao između regresionih linija će biti 90°.

****

U programu SPSS se dijagram raspršenja za dve varijable sa ucrtanom regresionom linijom može napraviti u meniju **Graphs/Scatter** tako da se u prozoru za dijalog **Scatterplot** klikne na tip **Simple** a potom na dugme **Define**. U prozoru **Simple Scatterplot** ime kriterijumske varijable se ubacuje u polje **Y Axes** a ime prediktorske varijable u polje **X Axes**. Po kliku na dugme **OK** u prozoru **Chart Carousel** (SPSS 6.1) ili prozoru za ispis (SPSS 8.0) dobija se skater dijagram. Da bi se u dijagram ucrtala regresiona prava za

regresiju **Y** na **X** varijablu treba kliknuti na dugme **Edit** i u meniju editora **Chart Carousel**-a (SPSS 6.1) ili dvaput na grafik (SPSS 8.0) i odabrati **Chart/Options**. U prozoru poddijaloga **Scatterplot Options** koji se otvori treba u okviru **Fit Line** uključiti dugme **Total** a zatim u poddijalogu **Scatterplot Options: Fit Line**, unutar okvira **Fit Method** uključiti opciju **Linear regression**, klikom na odgovarajuću ikonicu (ili najčešće samo proveriti da li je ova opcija uključena).
Za regresionu analizu sa jednom prediktorskom varijablom potrebno je odabrati meni **Statistics/Regression/Linear** i u dijalog prozoru **Linear Regression** u polje **Dependent** ubaciti ime kriterijumske a u polje **Independent(s)** ime prediktorske varijable. U poddijalogu **Statistics** može se odabrati željeni statistik. Od statistika treba uvek odabrati bar **Descriptives** (daje AS i SD varijabli i matricu interkorelacija varijabli), **Estimate** u okviru **Regression Coefficient** (ocene regresionih koeficijenata) i **Model Fit** (testiranje značajnosti koeficijenta multiple korelacije). U opciji **Save** određujemo šta treba sačuvati u aktivnom fajlu (poželjno je, mada nije uvek nužno, sačuvati nestandardizovane ili standardizovane predviđene vrednosti /**Unstandardized** ili **Standardized** u okviru **Predicted values**/ i nestandardizovane ili standardizovane reziduale/**Unstandardized** ili **Standardized** u okviru **Reziduals**/).
Na početku ispisa iz ove procedure dati su (posle AS, SD i matrice interkorelacija varijabli) koeficijent multiple korelacije (**Multiple R**), kvadrat ovog koeficijenta (**R Square**), korigovani koeficijent multiple determinacije (nepristrasna ocena populacijske vrednosti - **Adjusted R Square**) i standardna greška (**Standard error**) a potom u bloku **Analysis of Variance** test značajnosti koeficijenta linearne korelacije. Suma kvadrata za regresiju (**Regression** u koloni **Sum of Squares**), i suma kvadrata reziduala (**Residual**) podeljeni njima pripadajućim brojem stepeni slobode (za regresiju df=m a za reziduale df=n-m-1) daju prosečne kvadrate odstupanja (**Mean square**). Količnik prosečnog kvadrata za regresiju i prosečnog kvadrata za reziduale (grešku) je **F**-statistik. Vrednost **Signif F** je verovatnoća da se na slučajnom reprezentativnom uzorku date veličine dobije toliki F-statistik ili još veći ako je nulta hipoteza tačna. U bloku **Variables in the Equation** je data tabela sa nestandardizovanim (**B**) i standardizovanim regresionim koeficijentima (**BETA**). Količnik B/SE B daje **T** statistik, a u koloni **sig T** su rezultati testiranja značajnosti regresionih koeficijenata. **Constant** je intercept (odsečak), tj. vrednost kriterijumske varijable kada je vrednost prediktorske varijable jednaka nuli.

Primer:

Na uzorku izbeglica u našoj zemlji prikupljeni su, upitnikom za procenu ratnih stresora, podaci o ukupnoj izloženosti ratnim stresorima (borbenim,van linije fronta, zarobljeničkim i sl.). Ukupna relativna izloženost ratnim stresorima (variabla E) iskazana je kao proporcija u odnosu na broj od 50 reprezentativnih ratnih stresora. Osim toga, Mississipi skalom izmerena je izraženost PTSD simptoma. Da li se individualne razlike u izraženosti simptoma PTSD-a u izbegličkoj populaciji mogu predvideti ili objasniti linearnom vezom izrazenosti PTSD simptomatologije sa ekstenzitetom izloženosti ratnim stresorima? Ako mogu, koliko uspešno možemo da predvidimo intenzitet PTSD simptomatologije na osnovu ove prediktorske varijable?

Aritmetičke sredine, standardne devijacije i koeficijent linearne korelacije kriterijumske i prediktorske varijable:

	Mean	Std Dev	Label
MSUKUP	96.767	19.190	MISSISSIPPI SKALA-UKUPAN SKOR
E	.191	.151	RELATIVNA UKUPNA IZLOŽENOST STRESORIMA

N of Cases = 116

Correlation, 1-tailed Sig:

	MSUKUP	E
MSUKUP	1.000	.321
	.	.000
E	.321	1.000
	.000	.

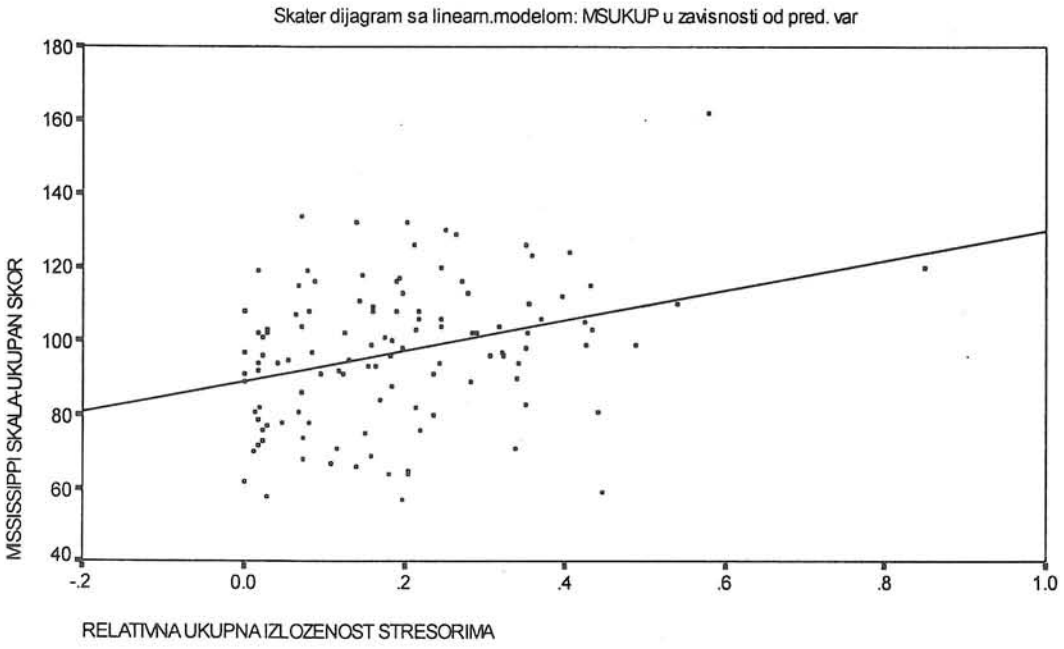
Test normalnosti zavisne, kriterijumske varijable MSUKUP u populaciji izbeglica kao jedan od preduslova za primenu modela jednostavne linearne regresije:

MSUKUP MISSISSIPPI SKALA-UKUPAN

Valid cases: 116.0 Missing cases: 92.0 Percent missing: 44.2

	Statistic	df	Significance
K-S (Lilliefors)	.0432	116	> .2000

Skaterdijagram sa regresionom pravom za regresiju MSUKUP-a na ukupnu izloženost stresorima:



```
          * * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Listwise Deletion of Missing Data

Equation Number 1    Dependent Variable..   MSUKUP   MISSISSIPPI SKALA-
UKUPAN SK

Block Number  1.  Method:  Enter      E

Variable(s) Entered on Step Number
   1..    E          RELATIVNA UKUPNA IZLOŽENOST STRESORIMA

Multiple R            .32096
R Square              .10301
Adjusted R Square     .09514
Standard Error      18.25411


Analysis of Variance
                    DF        Sum of Squares      Mean Square
Regression           1           4362.47496       4362.47496
Residual           114          37986.24055        333.21264

F =      13.09217       Signif F =  .0004

------------------ Variables in the Equation ------------------

Variable              B         SE B       Beta          T  Sig T

E             40.900098    11.303648    .320957      3.618  .0004
(Constant)    88.975285     2.740437                 32.468  .0000

End Block Number   1   All requested variables entered.

          * * * * * * * * * * * * * * * * * * * * * * * * * * * * *


Residuals Statistics:

             Min       Max      Mean   Std Dev    N

*PRED     88.9753 123.7404  96.7672    6.1591  116
*RESID   -48.2225  49.3746    .0000   18.1746  116
*ZPRED    -1.2651   4.3794    .0000    1.0000  116
*ZRESID   -2.6417   2.7048    .0000     .9956  116

Total Cases =      208

          * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

From Equation   1:   2 new variables have been created.

  Name        Contents
  ----        --------

  ZPR_1       Standardized Predicted Value
  ZRE_1       Standardized Residual
```

Test normalnosti reziduala

ZRE_1 Standardized Residual

Valid cases: 116.0 Missing cases: 92.0 Percent missing: 44.2

	Statistic	df	Significance
K-S (Lilliefors)	.0297	116	> .2000

Zaključak: Između izraženosti PTSD simptoma i ekstenziteta izloženosti stresorima postoji pozitivna linearna povezanost u izbegličkoj populaciji ali je ova povezanost umerenog intenziteta. Oko 10% ukupne varijanse u izraženosti PTSD simptoma možemo objasniti razlikama u ekstenzitetu izloženosti ratnim stresorima. Postavlja se pitanje da li bi naše objašnjenje (i predviđanje) bilo uspešnije kada bismo kao prediktore uzeli i određene dimenzije ličnosti? Odgovor na ovo pitanje dat je kroz primer u odeljku o multiploj regresiji.

****

XII. HI-KVADRAT TEST: TESTIRANJE HIPOTEZA O DISTRIBUCIJI VEROVATNOĆA KATEGORIČKE VARIJABLE I ANALIZA KONTINGENCIJSKIH TABELA

XII.1. Testiranje hipoteze o distribuciji verovatnoća kategoričke varijable

Struktura podataka:

$E = \{ e_i ; i = 1,...,n\} \subset P$ je slučajni uzorak izvučen iz populacije;
$Q = \{ q_j ; j = 1,2,..., r\}$ je kategorička varijabla koja se sastoji od r kategorija

Raspored verovatnoća po kategorijama kategoričke varijable opisuje stanje u populaciji. Iz poznavanja prirode date varijable (ili na osnovu određene teorijske pretpostavke) moguće je precizno matematički definisati ovaj raspored. Tako se, na primer, na osnovu određene (obično genetičke) teorije može izračunati verovatnoća da neki član populacije bude desnoruk, levoruk ili ambidekstar. Kada iz date populacije izaberemo slučajni reprezentativni uzorak, očekujemo da raspored dobijenih empirijskih frekvenci odstupa samo slučajno od očekivanog (teorijskog) rasporeda. Moguće je, pak, da je naša teorijska pretpostavka pogrešna i da se ova dva rasporeda razlikuju više no što bismo mogli očekivati samo na osnovu slučaja. Tada kažemo da je razlika empirijskog i očekivanog rasporeda statistički značajna.
Testiranje statističke značajnosti odstupanja rasporeda empirijskih frekvenci po kategorijama nominalne varijable od rasporeda očekivanog na osnovu nekog teorijskog modela svodi se na testiranje hipoteze:

$$H_o: p_k = \pi_k \ \forall k$$

(π_k su očekivane (teorijske) verovatnoće a p_k su dobijene, tj.empirijski ocenjene verovatnoće) ili, što je isto, hipoteze:

$$H_o: f_k = \phi_k \ \forall k$$

gde su f_k empirijski dobijene frekvence, a ϕ_k očekivane (u skladu sa nekom pretpostavkom) frekvence pri čemu je $\phi_k = \pi_k * n$.
Alternativna hipoteza je:

$$H_1: p_k \neq \pi_k \quad \text{ili:} \quad H_1: f_k \neq \phi_k \ \exists k$$

Test statistik za testiranje H_o je:

$$\chi^2 = \sum_{k=1}^{r} \frac{(f_k - \phi_k)^2}{\phi_k}$$

Test statistik χ^2 pod H_o ima distribuciju uzorkovanja koja je Hi-kvadrat funkcija gustine sa k-1 stepeni slobode. Računamo

$$P(\chi^2 \geq \chi^2_{dob} | H_0 \text{ tačno}) = \int_{\chi^2_{dob}}^{\infty} f(\chi^2) d\chi^2$$

tj. verovatnoću da χ^2 bude jednak ili veći od dobijenog pod pretpostavkom da je H_0 tačno. Ako je ova verovatnoća manja ili jednaka α odbacujemo H_0, a ako je veća od α ne odbacujemo H_0.

Zapamtite:

- Hi-kvadrat test podrazumeva da su frekvence **nezavisne**, tj. da se jedan entitet sme pojaviti samo jednom u jednoj kategoriji varijable, a nikako ponovo u istoj ili drugoj kategoriji.

- Opisani Hi-kvadrat test smemo koristiti ako su, u slučaju kada nominalna varijabla ima samo **dve** kategorije, **sve očekivane frekvence** bar 5 ili ako je reč o varijabli sa **više od dve** kategorije ukoliko očekivane frekvence **nisu manje od 5** u **20% kategorija** i **nijedna** očekivana frekvenca nije **manja od 1**. U situacijama /koje smo upravo naveli/ kada ovaj test ne treba koristititi moguće je primeniti tzv. **binomni test** ili, ako je to smisleno, treba **sažeti** kategorije na manji broj.

****

Testiranje hipoteze o distribuciji verovatnoća kategoričke varijable u SPSS-u:
Odabrati meni **Statistics/Nonparametric Tests/Chi-square**. Ime kategoričke varijable za koju testiramo hipotezu o verovatnoći ubacujemo u okvir **Test Variable list**. Podrazumevana distribucija je uniformna (očekivane frekvence jednake u svim kategorijama, tj. $\phi_k = n/k$). Ukoliko je reč o drugačijoj teorijskoj pretpostavci očekivane frekvence se moraju upisati u polje **Expected values** i to tako da se prvo unose očekivane frekvence kategorija koje su označene nižim ciframa (dakle, po rastućem redosledu cifara kojima su kategorije označene) . Posle svake unesene očekivane frekvence treba kliknuti na dugme **Add** da bi se ova vrednost upisala u prazan okvir u desnom uglu.

Primer: Prema genetičkom modelu rukosti M.Anett, rukost bi trebalo da ima binomnu distribuciju u populaciji tako da bi trebalo da bude 4% levorukih, 30% ambidekstara i 66% dešnjaka. U ovom modelu rukosti pretpostavlja se da je preferencija ruke uzrokovana prisustvom ili odsustvom genetskog faktora "pomeranja udesno".

Na slučajnom reprezentativnom uzorku od 413 ispitanika dobijeni su podaci o rukosti. Da li je empirijski dobijeni raspored rukosti u skladu sa genetičkim modelom M. Anett? Očekivane frekvence, na osnovu modela M. Anett, treba uneti u polje **Expected values** ovim redosledom: 16.5 [dobijamo iz (4/100)*413], 123.9 [dobijamo iz (30/100)*413], 272.6 [dobijamo iz (66/100)*413].

- - - - - Chi-Square Test

DOMINRUK dominantna ruka

Category	Cases Observed	Expected	Residual
leva	1	124	16.50	107.50
obe	2	75	123.90	-48.90
desna	3	214	272.60	-58.60

Total		413		

Chi-Square	D.F.	Significance
732.2753	2	.0000

Hi-kvadrat od 732.28 je, za dva stepena slobode, statistički značajan (significance je manje od 0.05). Raspored emprijski dobijenih frekvenci (kolona Cases Observed) statistički značajno odstupa od rasporeda frekvenci koje očekujemo na osnovu modela rukosti M.Anett. Uočimo i da je rezidual (razlika empirijskih i očekivanih frekvenci) najveći u kategoriji levorukih. Prema tome, dobijeni rezultati nisu u skladu sa modelom M. Anett.

****

XII.2.Analiza kontingencijskih tabela- testiranje nezavisnosti dveju kategoričkih varijabli

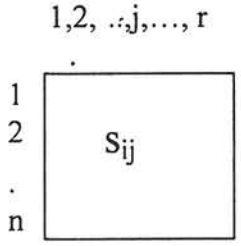
Struktura podataka:

$E = \{ e_i ; i = 1,...,n\} \subset P$ je slučajni uzorak izvučen iz populacije

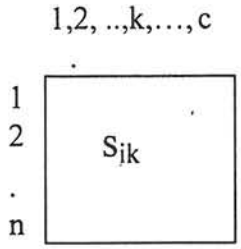
$Q_1 = \{ q_{1j} ; j = 1,2,..., r\}$ kategorička varijabla koja se sastoji od r kategorija

$Q_2 = \{ q_{2k} ; k = 1,2,..., c\}$ kategorička varijabla koja se sastoji od c kategorija

$E \otimes Q_1 = \mathbf{S_1}$



$E \otimes Q_2 = \mathbf{S_2}$



Ocene verovatnoća pojedinih kategorija Q_1 i Q_2 date su u dijagonalnim matricama:

$$\mathbf{P}_1 = \mathbf{F}_1 * n^{-1} \text{ i } \mathbf{P}_2 = \mathbf{F}_2 * n^{-1}$$

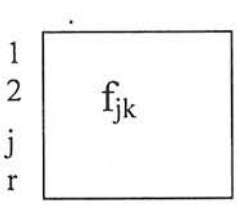
gde su matrice $\mathbf{F}_1$ i $\mathbf{F}_2$ definisane kao:

$$\mathbf{F}_1 = \mathbf{S}_1{}^t\mathbf{S}_1 \text{ i } \mathbf{F}_2 = \mathbf{S}_2{}^t\mathbf{S}_2$$

U glavnoj dijagonali matrica $\mathbf{F}_1$ i $\mathbf{F}_2$ su marginalne frekvence po kategorijama varijabli Q_1 i Q_2.

Kontingencijska tabela ili tabela unakrsnog razvrstavanja definisana je kao:

$$\mathbf{F} = \mathbf{S}_1{}^t\mathbf{S}_2$$



Elementi matrice $\mathbf{F}$ su zajedničke frekvence jer je f_{jk}= numerus ($e_i \in q_j \wedge e_i \in q_k$). Ocene zajedničkih verovatnoća [$p_{jk} = P(e_i \in q_j \cap q_k)$] su date u matrici:

$$\mathbf{P} = \mathbf{F} * n^{-1}$$

Marginalni rasporedi su zbirovi redova ili kolona koji se dobijaju sabiranjem zajedničkih frekvenci kontingencijske tabele.

Problem: Postoji li povezanost dveju kategoričkih varijabli ili su one međusobno nezavisne?

Nulta hipoteza:

H_o: $Q_1 \perp Q_2$

(Varijable Q_1 i Q_2 su međusobno nezavisne)

ili

H_o: $p_{jk} = p_j * p_k$; $\forall j$ i $\forall k$

(Relativne zajedničke frekvence u svim ćelijama jednake su proizvodu odgovarajućih relativnih marginalnih frekvenci)

ili

H_o: $f_{jk} = \phi_{jk}$; $\forall j$ i $\forall k$

(Zajedničke empirijske frekvence jednake su očekivanim frekvencama za svako j i svako k)

Alternativna hipoteza je:

H_1: $p_{jk} \neq p_j * p_k$

ili

H_1: $f_{jk} \neq \phi_{jk}$

f_{jk} su dobijene, empirijske zajedničke frekvence u ćelijama kontingencijske tabele, a ϕ_{jk} su očekivane zajedničke frekvence pod pretpostavkom da su kategoričke varijable nezavisne: $\phi_{jk}= (f_j * f_k)/n$ gde su f_j i f_k zbirovi redova i kolona kojima ćelija pripada.[81]
Razliku očekivane i empirijske frekvence u datoj ćeliji zovemo i rezidualom.

Test statistik :

$$\chi^2 = \sum_{k=1}^{c} \sum_{j=1}^{r} \frac{(f_{jk} - \phi_{jk})^2}{\phi_{jk}}$$

pod Ho ima Hi-kvadrat distribuciju uzorkovanja sa (r-1) (c-1) stepeni slobode.[82] Potom računamo $P(\chi^2 \geq \chi^2_{\text{dobijeni}} \mid \text{Ho tačno})$. Ako je ova verovatnoća manja ili jednaka α odbacujemo Ho, ako je veća ne odbacujemo Ho.

Zapamtite:

- Opisani Hi-kvadrat test smemo koristiti ako su, u slučaju kada nominalne varijable imaju po **dve** kategorije sve **očekivane** frekvence bar **5** ili, ako je reč o varijablama sa **više od dve** kategorije, ukoliko očekivane frekvence nisu manje od 5 u **20% kategorija** i **nijedna** očekivana frekvenca **nije manja od 1**. U situacijama kada ovaj test ne treba koristititi, a koje smo upravo naveli, treba, ako je to smisleno, sažeti kategorije na manji broj ili:
  Ako je n>40 a tabela 2x2 primeniti Yates-ovu korekciju[83]
  Ako je n < 40 a tabela 2x2 primeniti Fisher-ov egzaktni test.

- Prikazani postupak testiranja hipoteze o nezavisnosti podrazumeva da je uzorak *E* izvučen slučajno iz populacije *P*, tj. da je pri tom fiksirano samo n, tj. veličina uzorka. Identičan postupak se može primeniti i za testiranje hipoteza o interakciji ili homogenosti. **Hipoteza o interakciji** podrazumeva, pored **fiksiranog n** i **fiksirane marginalne verovatnoće**, dok **hipoteza o homogenosti** podrazumeva da je, pri izboru uzorka, **fiksirano n** i **marginalna distribucija** verovatnoća **jedne** od varijabli. Tako, na primer, ako smo pri izboru uzorka odlučili da u uzorak bude uključeno podjednako žena i muškaraca marginalna distribucija varijable pol je u tom slučaju fiksirana. U tom primeru smemo testirati hipotezu homogenosti polova u pogledu neke druge kategoričke varijable, kao i hipotezu o distribuciji verovatnoća te druge kategoričke varijable, ali ne i hipotezu o distribuciji verovatnoća za varijablu pol. Test nezavisnosti dopušta i testiranje hipoteza o raspodelama svake kategoričke varijable ponaosob. Test homogenosti dozvoljava i testiranje hipoteze o raspodeli za varijablu čija marginalna raspodela nije fiksirana dok test interakcije podrazumeva samo testiranje hipoteze o zajedničkim frekvencama.

---

[81] Ako su kategoričke varijable nezavisne (a pod tom pretpostavkom računamo zajedničke očekivane frekvence) onda je verovatnoća zajedničkog događanja neke kategorije varijable Q_1 i (istovremeno) neke kategorije varijable Q_2 jednaka proizvodima verovatnoće događanja date kategorije Q_1 varijable i verovatnoće događanja date kategorije varijable Q_2. Dakle:
$P[(Q_1=q_j) \wedge (Q_2=q_k)] = P(Q_1=q_j) * P(Q_2=q_k)$.
Tada $p_{jk} = p_j * p_k = (f_j/n) * (f_k/n)$, a $\phi_{jk} = p_{jk} * n = (f_j/n) * (f_k/n) * n = (f_j * f_k)/n$.

[82] Broj stepeni slobode je ovoliki jer moramo oceniti (r-1) verovatnoća za Q_1 i (c-1) verovatnoća za Q_2. Verovatnoće poslednje kategorije svake varijable determinisane su činjenicom da je

$$\sum_{j=1}^{r} p_j = 1 \text{ i } \sum_{k=1}^{c} p_k = 1.$$

[83] O Jejtsovoj korekciji se može pročitati u : Č.Dragićević Statistika za psihologe, str. 230 i 234.

XII.3.Koeficijenti asocijacije dveju kategoričkih varijabli

U slučaju odbacivanja nulte hipoteze smisaono je upotrebiti neku od mera intenziteta asocijacije kategoričkih varijabli[84]:

1. **Kramerov V koeficijent**:

$$V = \sqrt{\chi^2 / [n(q-1)]} \text{ gde je } q = \min(r, c)$$

Gornja granica ovog koeficijenta je 1. Kramerov V-koeficijent, međutim, ima tendenciju da potcenjuje stvarnu jačinu povezanosti.

U slučaju kada je tabela 2 x 2 ovaj je koeficijent identičan Fi-koeficijentu :

$$\phi = \sqrt{\chi^2 / n}$$

Za tabele 2x2 smisaono je upotrebiti i **količnik šansi** (KŠ) ili **količnik unakrsnih proizvoda** (engl. odds ratio). Ako su f_{11}, f_{12} , f_{22} i f_{21} empirijski dobijene frekvence u ćelijama ove tabele onda je $KŠ = (f_{11} / f_{12}) / (f_{21} / f_{22}) = (f_{11} * f_{22}) / (f_{12} * f_{21})$.
Ukoliko su kategoričke varijable (svaka sa po dve kategorije) statistički nezavisne KŠ je jednak 1. Inače, ovaj količnik varira od 0 do ∞. Međutim, on nije simetričan oko 1 pa može biti različit za tabele u kojima je prisutan isti stepen asocijacije ali drugačijeg smera. Da bi bio simetričan oko 1 dovojno je uzeti logaritam od KŠ. Log KŠ varira od -∞ do ∞ a kada su varijable nezavisne jednak je 0. Količnik šansi se može generalizovati i na tabele kontingencije veće od 2x2.

2. **Pirsonov koeficijent kontingencije** :

$$C = \sqrt{\chi^2 / (n + \chi^2)}$$

C-koeficijent ne dostiže vrednost 1, a maksimalna vrednost koju može dostići zavisi od dimenzija tabele:

$$C_{max} = \sqrt{q - 1 / q} \text{ gde je } q = \min(r,c)$$

Jasno je stoga da se C-koeficijenti iz tabela kontingencije koje su nejednakih dimenzija ne mogu direktno porediti.

---

[84] Podroban prikaz koeficijenta asocijacije može se pročitati u Momirović (1988).

3. **Lambda koeficijent**:

3a. Asimetrični oblik:

$$\lambda_{r(\text{kolone nezavisne})} = \frac{\sum_{j=1}^{r} \max_k f_{jk} - \max_k f_k}{n - \max_k f_k}$$

$$\lambda_{c(\text{redovi nezavisni})} = \frac{\sum_{k=1}^{c} \max_j f_{jk} - \max_j f_j}{n - \max_j f_j}$$

Asimetrična Lambda pokazuje proporcionalno smanjenje verovatnoće greške u predviđanju kategorijalne pripadnosti ispitanika na "zavisnoj" kategoričkoj varijabli na osnovu poznavanja kategorijalne pripadnosti na "nezavisnoj" varijabli u odnosu na verovatnoću greške kada tu informaciju nemamo.

3b. Simetrični oblik:

$$\lambda = \frac{\sum_{j=1}^{r} \max_k f_{jk} + \sum_{k=1}^{c} \max_j f_{jk} - \max_k f_k - \max_j f_j}{2n - \max_k f_k - \max_j f_j}$$

4. **Goodman i Kruskalov tau koeficijent** takođe govori o relativnoj redukciji greške u predviđanju kategorijalne pripadnosti na jednoj varijabli na osnovu poznatavanja kategorijalne pripadnosti u drugoj varijabli:

4a. Asimetrični oblik:

Za predviđanje Q_1 na osnovu Q_2:

$$\tau_{1.2} = \frac{\sum_{j=1}^{r} \sum_{k=1}^{c} [(n * f_{jk} - f_j * f_k)^2 / f_j]}{n(n^2 - \sum_{k=1}^{c} f_k^2)}$$

Za predviđanje Q_2 na osnovu poznavanja Q_1:

$$\tau_{2.1} = \frac{\sum_{j=1}^{r} \sum_{k=1}^{c} [(n * f_{jk} - f_j * f_k)^2 / f_k]}{n(n^2 - \sum_{j=1}^{r} f_j^2)}$$

Kada je $r = c = 2$ onda je $\tau_{1.2} = \tau_{2.1}$

4b) Simetrični oblik:

$$\tau = \frac{\sum_{j=1}^{r}\sum_{k=1}^{c}[(n * f_{jk} - f_j * f_k)^2 / (f_j + f_k)]}{n(2 * n^2 - \sum_{j=1}^{r} f_j^2 - \sum_{k=1}^{c} f_k^2)}$$

5. Koeficijent neizvesnosti:

5a. Asimetrični oblik:

$$KN(Q_1 \text{ nezavisna}) = \frac{H(Q_2) - H(Q_2|Q_1)}{H(Q_2)} \text{ gde je } H(Q_2|Q_1) = -\sum_{j=1}^{r}\sum_{k=1}^{c} p_{jk} \log p_{(k|j)}$$

$$KN\ (Q_2 \text{ nezavisna}) = \frac{H(Q_1) - H(Q_1|Q_2)}{H(Q_1)} \text{ gde je } H(Q_1|Q_2) = -\sum_{k=1}^{c}\sum_{j=1}^{r} p_{jk} \log p_{(j|k)}$$

5b. Simetrični oblik:

$$KN = \frac{H(Q_1) + H(Q_2) - H(Q_1 \cap Q_2)}{\min[H(Q_1), H(Q_2]}$$

gde je prosečna zajednička neizvesnost :

$$H(Q_1 \cap Q_2) = -\sum_{j=1}^{r}\sum_{k=1}^{c} p_{jk} \log p_{jk}$$

****

Hi-kvadrat test nezavisnosti dveju kategoričkih varijabli nalazi se u meniju **Statistics/Summerize/Crosstabs**. U dijalog prozoru jednu kategoričku varijablu treba definisati kao **Raw** variable a drugu kao **Column** Variable. Potom u opciji **Statistics** odabrati željeni statistik klikom na kvadratić ispred naziva statistika (poželjno je odabrati bar **Chi-square** i neke od mera asocijacije iz okvira **Nominal** Data, npr.**Cramer's V**). U opciji **Cells** može se definisati željeni sdržaj svake ćelije tabele na ispisu (procenti -okvir **Percentages** /u odnosu na zbir reda (**Raw**), zbir kolone (**Column**) i u odnosu na ukupno n (**Total**)/, očekivane frekvence (**Counts Expected**), sirovi reziduali /razlike empirijskih i očekivanih frekvenci- Okvir **Residuals** opcija **Unstandardized**/, standardizovani reziduali /razlike očekivanih i empirijskih frekvenci podeljene korenom iz očekivane frekvence- **Standardized**/ i korigovani standardizovani reziduali / reziduali podeljeni ocenom njihove standardne greške- **Adj. standardized**/.

U ispisu iz ove procedure dobija se kontingencijska tabela sa definisanim sadržajem ćelija (svakako sadrži empirijske frekvence, a ostali elementi ako su definisani opcijom Cells dijaloga Crosstabs-šta svaka ćelija u tabeli sadrži napisano je u gornjem levom uglu tabele). Ispod tabele je Pearsonov Hi-kvadrat statistik(**Pearson**), broj stepeni slobode(**df**) i verovatnoća da Hi-kvadrat statistik bude jednak ili veći od dobijenog ako je Ho tačno (**Significance**). Zatim je data najmanja očekivana frekvenca (**Minimum expected frequency**) i broj (i procenat) ćelija sa očekivanim frekvencama manjim od 5 (**Cells with expexted frequencies < 5**). Za tabele 2x2 daje se rezultat sa i bez Yates- ove korekcije, a, ukoliko je ukupan broj slučajeva manji od 20, umesto Hi-kvadrat testa dati su rezultati

Fisher-ovog testa egzaktne verovatnoće. Osim Pirsonovog Hi-kvadrat statistika u ispisu se pojavljuju i Hi-kvadrat količnika verodostojnosti (Likelihood Ratio) i Mantel-Haenszel test linearne asocijacije. Test linearne asocijacije podrazumeva da su kategorije bar sa ordinalnog nivoa merenja. Ukoliko su u opciji Statistics odabrani u ispisu su dati i koeficijenti asocijacije dveju kategoričkih varijabli (Koeficijent kontingencije, Fi-koeficijent, Kramerov V-koeficijent, Gudman-Kraskalov tau- koeficijent, lambda-koeficjent i koeficijent neizvesnosti).

Primer 1: Postoji li povezanost između preferencije ruke i dominantnosti oka? Kao što je poznato, kod svih parnih organa ljudskog organizma postoji tendencija da jedan od njih dominira. Postavlja se pitanje da li isti (genetski ili neki drugi) faktori determinišu lateralnost dominacije različitih organa. Ako je to slučaj onda bismo očekivali da dominantnost jednog od parnih organa (npr. ruke) bude povezana sa lateralnošću dominantnog člana para i za druge organe (npr. oka).
Na slučajnom uzorku od 413 ispitanika ispitana je rukost i dominantnost oka.[85]

Dobijeni rezultati su dati u Tabeli:

DOMINRUK dominantna ruka by DOMINOKO

		DOMINOKO			Page 1 of 1
	Count / Exp Val / Row Pct / Col Pct / Tot Pct	1	2	3	Row Total
DOMINRUK					
1 leva		35	62	27	124
		36.0	58.5	29.4	30.0%
		28.2%	50.0%	21.8%	
		29.2%	31.8%	27.6%	
		8.5%	15.0%	6.5%	
2 obe		27	28	20	75
		21.8	35.4	17.8	18.2%
		36.0%	37.3%	26.7%	
		22.5%	14.4%	20.4%	
		6.5%	6.8%	4.8%	
3 desna		58	105	51	214
		62.2	101.0	50.8	51.8%
		27.1%	49.1%	23.8%	
		48.3%	53.8%	52.0%	
		14.0%	25.4%	12.3%	
Column Total		120 29.1%	195 47.2%	98 23.7%	413 100.0%

Chi-Square	Value	DF	Significance
Pearson	3.93844	4	.41440
Likelihood Ratio	3.96689	4	.41051
Mantel-Haenszel test for linear association	.20036	1	.65443
Minimum Expected Frequency -	17.797		

Statistic	Value	ASE1	Val/ASE0	Approximate Significance
Phi	.09765			.41440 *1
Cramer's V	.06905			.41440 *1

[85] Kao što postoje različiti testovi kojima se utvrđuje rukost, tj. dominantno korišćenje jedne ruke u obavljanju složenih motornih veština, tako postoje i specijalne probe kojima se može odrediti na koje oko se ispitanik dominantno oslanja u opažanju.

Contingency Coefficient	.09719			.41440 *1
Lambda :				
symmetric	.00000	.00000		
with DOMINRUK dependent	.00000	.00000		
with DOMINOKO dependent	.00000	.00000		
Goodman & Kruskal Tau :				
with DOMINRUK dependent	.00364	.00371		.55848 *2
with DOMINOKO dependent	.00558	.00553		.33128 *2
Uncertainty Coefficient :				
symmetric	.00465	.00464	1.00225	.41051 *3
with DOMINRUK dependent	.00475	.00473	1.00225	.41051 *3
with DOMINOKO dependent	.00455	.00454	1.00225	.41051 *3

*1 Pearson chi-square probability
*2 Based on chi-square approximation
*3 Likelihood ratio chi-square probability
Number of Missing Observations: 0

U svakoj ćeliji Tabele prvi broj predstavlja (što je inače objašnjeno u gornjem levom uglu tabele) empirijsku frekvencu, drugi broj je očekivana frekvenca a ispod toga su procenat u odnosu na zbir reda kojem ćelija pripada, procenat u odnosu na zbir kolone u kojoj je ćelija i procenat u odnosu na ukupan broj slučajeva). Hi-kvadrat za testiranje hipoteze o nezavisnosti dveju kategoričkih varijabli je 3.94 (u redu **Pearson** kolone **Chi-square**) . Hi-kvadrat za 4 stepena slobode nije statistički značajan (α = .41440) i, prema tome, nemamo razloga da odbacimo nultu hipoptezu. Možemo, dakle, zaključiti da između rukosti i dominantnosti oka nema povezanosti u populaciji. Budući da Hi-kvadrat nije statistički značajan nikakvo dalje tumačenje dobijenih mera asocijacije nema smisla.

Primer2: Postoji li povezanost između polne pripadnosti ispitanika i načina odgovaranja na stavku iz Gilford-Zimermanovog pregleda temperamenta: "Kada ugledate velike bube i pauke, podiđe Vas jeza"? Drugačije rečeno, da li se muškarci i žene u studentskoj populaciji razlikuju po načinu odgovaranja na ovoj stavci?

POL by G255

		G255		Page 1 of 1
	Count / Exp Val / Row Pct / Col Pct / Tot Pct	ne	da	Row Total
POL				
muski	0	301 235.0 62.7% 64.0% 31.4%	179 245.0 37.3% 36.5% 18.6%	480 50.0%
zenski	1	169 235.0 35.2% 36.0% 17.6%	311 245.0 64.8% 63.5% 32.4%	480 50.0%
	Column Total	470 49.0%	490 51.0%	960 100.0%

Chi-Square	Value	DF	Significance
Pearson	72.63152	1	.00000
Continuity Correction	71.53521	1	.00000
Likelihood Ratio	73.57803	1	.00000
Mantel-Haenszel test for linear association	72.55587	1	.00000

Minimum Expected Frequency - 235.000

Statistic		Value	ASE1	Val/ASE0	Approximate Significance
Phi		-.27506			.00000 *1
Cramer's V		.27506			.00000 *1
Contingency Coefficient		.26521			.00000 *1
Lambda :					
symmetric		.26737	.03480	7.15338	
with POL	dependent	.27500	.03927	6.07676	
with G255	dependent	.25957	.04011	5.66068	
Goodman & Kruskal Tau :					
with POL	dependent	.07566	.01707		.00000 *2
with G255	dependent	.07566	.01707		.00000 *2
Uncertainty Coefficient :					
symmetric		.05530	.01264	4.37435	.00000 *3
with POL	dependent	.05529	.01264	4.37435	.00000 *3
with G255	dependent	.05530	.01264	4.37435	.00000 *3

*1 Pearson chi-square probability
*2 Based on chi-square approximation
*3 Likelihood ratio chi-square probability

Statistic	Value	95% Confidence Bounds	
Relative Risk Estimate (POL 0 / POL 1) :			
case control	3.09448	2.37824	4.02642
cohort (G255 0 Risk)	1.78107	1.54901	2.04788
cohort (G255 1 Risk)	.57556	.50366	.65773

Number of Missing Observations: 0

Hi-kvadrat (72.63) je statistički značajan i na osnovu toga odbacujemo hipotezu o homogenosti.[86] Studenti muškarci različito odgovaraju na ovu stavku od studentkinja. Pregledom tabele uočavamo da 64.8% studentkinja daje potvrdan odgovor dok 37.3% studenata muškog pola odgovara potvrdno na ovu stavku. Ipak, ova razlika nije tako velika kao što bi se možda očekivalo. Kramerov V koeficijent iznosi (samo) 0.275. Na osnovu lambde i koeficijenta neizvesnosti uočavamo da je asocijacija između pola i načina odgovaranja na ovu stavku praktično simetrična: poznavanje pola nam pomaže da predvidimo način odgovaranja isto onoliko koliko nam i informacija o načinu odgovaranja pomaže da pogodimo kojeg je pola ispitanik.

Iz ocena relativnog rizika /**Relative Risk Estimate** (POL 1/POL 2)/ uočavamo da je relativni rizik potvrdnog odgovora duplo manji za muškarce no za žene (179/311=0.58) dok je "rizik" odrečnog odgovora za 1.78 puta veći za muškarce u odnosu na žene (301/169=1.78).

Količnik šansi (**odds ratio**) je (301/179)/(169/311) = 3.09. Ovo govori da je šansa da se odgovori odrečno (u odnosu na šansu potvrdnog odgovora) tri puta veća ako je reč o muškarcima nego ako je reč o ženama.

****

[86] Budući da je uzorak biran slučajno ali tako da u uzorku bude podjednak broj studenata oba pola ovde ima smisla testirati hipotezu homogenosti, a ne nezavisnosti. Dakle, marginalna raspodela varijable pol je fiksirana u ovom slučaju.

XIII. NEPARAMETRIJSKE STATISTIČKE METODE (METODE BEZ STROGIH PRETPOSTAVKI O OBLIKU DISTRIBUCIJE MERA U POPULACIJI) ZA RANGOVANE PODATKE

XIII.1. Parametrijske i "neparametrijske" metode

Tehnike statističkog zaključivanja, tj. statistički testovi koje koristimo da bismo na osnovu podataka dobijenih istraživanjem doneli odgovarajuće zaključke, zasnivaju se na određenim statističkim modelima. Statistički model koji je u osnovi nekog statističkog testa obuhvata određene pretpostavke o prirodi raspodele varijabli u populaciji i načinu izvlačenja uzorka iz populacije. Ove pretpostavke su uslovi za primenu određenog statističkog postupka. Pored toga, svaki statistički test podrazumeva određeni tip merne skale sa koje dolaze brojevi. Statistički testovi koji imaju najstrože i najmnogobrojnije zahteve u pogledu nivoa merenja i prirode distribucije varijabli u populaciji su ujedno i testovi sa najvećom *snagom*.[87]

Parametrijski statistički testovi počivaju na modelu koji postavlja određene pretpostavke o distribucijama varijabli u populaciji iz koje je izvučen uzorak. Kada su ove pretpostavke o distribuciji populacije napravljene, populacija se može izraziti u terminima parametara. Ako se, na primer, pretpostavi da je varijabla normalno distribuirana njena distribucija se može izraziti u terminima μ i σ. Na taj način se problem statističkog zaključivanja svodi na zaključivanje o parametrima a metode koje se pri tome koriste su parametrijske metode. Osnovni i najpoznatiji uslov većine parametrijskih testova je normalna raspodela mera u populaciji. Na primer, t-test i analiza varijanse zasnovani su na pretpostavkama da se mere zavisne varijable normalno raspodeljuju u populaciji i da su varijanse subpopulacija definisanih nivoima faktora (nezavisne varijable) homogene. Mada se to obično izostavlja, zaključak koji proističe iz ovih postupaka trebalo bi, zapravo, da počinje sa: "Ako su pretpostavke o obliku distribucije mera u populaciji valjane onda...". Smisaonost rezultata parametrijskih testova zavisi od valjanosti ovih pretpostavki. Pored toga većina parametrijskih testova zahteva podatke bar sa intervalne merne skale.

Često, međutim, o obliku distribucije mera u populaciji ništa ne znamo, a katkad znamo da ne možemo pretpostaviti normalnu raspodelu (npr. kod socijalnih stavova). Isto tako, podaci ne mogu uvek poticati sa intervalne ili racio merne skale pa ne dolazi u obzir primena parametrijskih statističkih testova.

Neparametrijski statistički testovi ne postavljaju stroge uslove u pogledu oblika distribucije mera u populaciji, tj. slobodni su od pretpostavke o obliku raspodele mera u populaciji (engl. distribution free tests).[88] Zaključci na osnovu neparametrijskih tehnika mogu početi

---

[87] Snaga statističkog testa, kao što je poznato, iskazuje se verovatnoćom odbacivanja pogrešne nulte hipoteze. Prema tome, snaga statističkog testa je 1- β, pri čemu je β verovatnoća II tipa greške u statističkom zaključivanju, tj. verovatnoća da pogrešimo **ne odbacujući** nultu hipotezu ako je ona pogrešna. Snaga statističkog testa raste sa povećanjem veličine uzorka.

[88] O još nekim neparametrijskim statističkim testovima koji nisu obrađeni u Priručniku može se pročitati u Vuković (1997), str. 435 - 526.

sa : "Bez obzira na oblik distribucije mera u populaciji...". Ovi testovi ne postavljaju nikakve zahteve u pogledu parametara populacije iz koje su uzorci izvučeni, a uslovi koje zahtevaju su daleko blaži od onih kod parametrijskih testova (npr. nezavisnost opservacija i kontinuiranost varijabli), tako da možemo reći da su slobodni, tj. nezavisni od pojedinih karakteristika distribucije mera varijable u populaciji. Pored toga, oni se mogu primenjivati i na podacima sa ordinalne ili čak sa nominalne merne skale.

Prednosti neparametrijskih statističkih testova

1. Ne postavljaju zahteve o obliku distribucije mera u populaciji pa se mogu primeniti u velikom broju situacija.
2. Mogu se primeniti za poređenje uzoraka iz različitih (po obliku distribucije mera) populacija.
3. Mogu se primeniti na rangove i na nominalne podatke.
4. Mogu se primeniti i na veoma malim uzorcima (manjim od 10).
5. Lakši su za učenje i primenu.

Nepovoljnost neparametrijskih statističkih testova

Ako su ispunjeni svi uslovi za primenu parametrijskog statističkog modela primenom neprametrijskog testa gube se neke informacije sadržane u podacima. Tako, npr. ako su ispunjeni uslovi za primenu t-testa za nezavisne uzorke, u kojem slučaju su podaci za kvantitativnu varijablu sa intervalne ili racio skale, rangovanjem ovih podataka da bi se primenio neki od neparametrijskih postupaka (npr. Man-Vitnijev test) gube se informacije o tome **za koliko jedinica** se ispitanici razlikuju na toj varijabli, a ostaje samo informacija o njihovom redosledu na toj varijabli.

XIII.2.Neparametrijske metode korelacije rangova

Spirmanov koeficijent korelacije rangova (r_S)

Namena: Za određivanje slaganja, tj. monotone povezanosti dva niza rangova.

Uslovi: Podaci moraju poticati bar sa ordinalne merne skale.

Postupak:

-Za svaki par rangova računaju se diferencije (d_i) a potom razlike između parova rangova kvadriraju

-Kvadrirane razlika rangova se saberu i dobija $\sum_{i=1}^{n} d_i^2$

-Računa se vrednost izraza:[89]

$$r_S = 1 - \frac{2\sum_{i=1}^{n} d_i^2}{\max \sum_{i=1}^{n} d_i^2}$$

-Zamenom $\max \sum_{i=1}^{n} d_i^2$ sa $[n(n^2-1)]/3$, gde n predstavlja broj parova rangova, dobijamo:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Testiranje statističke značajnosti r_S

Za testiranje: H_0: $\rho = 0$, kada je $n \geq 10$ računamo statistik:

$$t = r_S \sqrt{\frac{n-2}{1-r_S^2}}$$

Test statistik t ima pod H_0 Studentovu distribuciju uzorkovanja sa n-2 stepeni slobode. Ako je $P(|t| \geq t_{dobijeno}) \leq 0.05$ odbacujemo H_0. Ako je $n < 10$ potrebno je konsultovati posebne tablice.[90]

Kendalov tau koeficijent korelacije rangova (r_K)

Namena i uslovi: Isti kao za Spirmanov koeficijent korelacije rangova.

Postupak:

- Računamo izraz: $r_K = \frac{S}{\frac{n(n-1)}{2}} = \frac{S}{S_{max.}}$

u kojem je n broj parova rangova, S pokazatelj uređenosti niza rangova, a

$$S_{max} = \frac{n(n-1)}{2}$$

---

[89] Očigledno je da:

-Kada su rangovi u inverznom redosledu $\sum_{i=1}^{n} d_i^2 = \max \sum_{i=1}^{n} d_i^2$, pa je $r_S = -1$.

-Ukoliko su nizovi rangova nezavisni $2\sum_{i=1}^{n} d_i^2 = \max \sum_{i=1}^{n} d_i^2$, pa je $r_S = 0$.

-Kada je slaganje dva niza rangova potpuno $\sum_{i=1}^{n} d_i^2 = 0$, pa je $r_S = 1$.

[90] Tablice egzaktnih distribucija statistika koje se spominju u ovom tekstu mogu se naći u udžbenicima neparametrijske statistike

- Vrednost izraza S se određuje tako što se rangovi u jednom nizu (rangovi jedne varijable) postave u prirodni redosled (1, 2, 3...), a potom se posmatraju svi mogući parovi rangova u drugom nizu (rangovi druge varijable). Broj takvih parova jednak je broju kombinacija od po 2 elementa iz skupa od n elemenata, tj. n(n -1)/2. (O kombinacijama se može pročitati u Dodatku 2, u odeljku I.2. Osnovni pojmovi i principi kombinatorike). Za svaki mogući par rangova u drugom nizu u kojem su rangovi u prirodnom redosledu daje se ocena +1, a za svaki par rangova koji nije u prirodnom redosledu ocena -1. Vrednost izraza S je zbir pozitivnih i negativnih jedinica.

Primer:

I niz rangova:1 2 3 4 5

II niz rangova:1 4 3 5 2

Par 1-4 dobija ocenu +1; 1-3 ocenu +1 1-5 ocenu +1;1-2 ocenu +1; 4-3 dobija -1; 4-5 ocenu +1; 4-2 ocenu -1; 3-5 ocenu +1;3-2 ocenu -1 i 5-2 ocenu -1. Vrednost izraza S u ovom primeru je +6-4=2.

Testiranje značajnosti r_K koeficijenta

Za testiranje: H_0: $\rho = 0$, kada je $n \geq 10$ računamo statistik:

$$z = (r_K - \mu_{rK})/ S_k$$

Statistik z ima pod H_0 normalnu distribuciju uzorkovanja sa sa $\mu_{rK} = 0$ i

$$S_k = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

Ako je $P(|z| \geq z_{dobijeno}) \leq 0.05$ odbacujemo H_0.

Kada je $n < 10$ potrebno je izračunati egzaktnu verovatnoću statistika pod H_0 ili je očitati iz posebnih tablica koje se nalaze u udžbenicima neparametrijske statistike.

Za iste podatke r_K je obično niži od r_S koeficijenta. Ipak, zaključci o značajnosti su isti. Prednost r_K koeficijenta nad r_S koeficijentom je u tome što se za r_K koeficijent može računati i parcijalna korelacija.

Zapamtite: Podaci koji su dobijeni sa intervalne ili racio skale mogu se (ali i ne moraju) pre pozivanja menija za računanje Spirmanovog i Kendalovog koeficijenta pretvoriti u rangove u meniju **Transform/Rank Cases**. Ukoliko se između dva niza rangovanih podataka izračuna Pirsonov produkt-moment korelacije on će biti jednak Spirmanovom koeficijentu korelacije rangova.

****

Spirmanov i Kendalov koeficijent korelacije dva niza rangova se u programu SPSS dobijaju menijem **Statistics/Correlate/Bivariate** tako što se u dijalogu **Bivariate Correlations** u polje **Variables** unesu imena varijabli u kojima su nizovi rangova i uključe opcije **Kendall's tau-b** i **Spearman** u okviru **Correlation Coefficients**. U ispisu su posle naziva ovih koeficijenata date njihove vrednosti, u zagradi broj parova rangova (**N**) a pod nazivom **Sig** verovatnoća greške I tipa pri odbacivanju nulte hipoteze.

Primer: Učenici su od strane nastavnika rangovani u pogledu disciplinovanosti na časovima matematike, a zatim su rangovani i u pogledu uspeha na testu znanja matematike.Postoji li povezanost između disciplinovanosti na časovima i uspeha?

```
K E N D A L L   C O R R E L A T I O N   C O E F F I C I E N T S

USPEH            .7200
              N(   25)
              Sig .000

               DISCIPL

(Coefficient / (Cases) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
S P E A R M A N   C O R R E L A T I O N   C O E F F I C I E N T S

USPEH            .8923
              N(   25)
              Sig .000

               DISCIPL

(Coefficient / (Cases) / 2-tailed Significance)

- -  Correlation Coefficients  - -

              DISCIPL     USPEH

DISCIPL       1.0000       .8923
              (   25)     (   25)
              P= .        P= .000

USPEH          .8923      1.0000
              (   25)     (   25)
              P= .000     P= .

(Coefficient / (Cases) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Uočimo da je Kendalov koeficijent aritmetički niži od Spirmanovog ali je zaključak o njihovoj značajnosti isti: oba koeficijenta su statistički značajna (Sig .000). Isto tako, pod naslovom - - **Correlation Coefficients** - - dat je Pirsonov koeficijent linearne korelacije izračunat na osnovu dva niza rangovanih podataka: na taj način izračunat ovaj koeficijent je matematički identičan Spirmanovom koeficijentu. Dakle, između disciplinovanosti na časovima matematike i uspeha u matematici postoji pozitivna monotona povezanost u učeničkoj populaciji.

****

Kendalov koeficijent slaganja (konkordance) (W)

Namena: Za određivanje slaganja, tj povezanosti između više od dva niza rangova.

Uslovi: Podaci moraju poticati bar sa ordinalne merne skale. Podaci su uređemi tako da su kriterijumi rangovanja (ili procenjivači) u kolonama, a entiteti koji su rangovani prema datim kriterijumima ili od strane nekoliko nezavisnih procenjivača nalaze se u redovima.

Postupak:

-Za svaki entitet koji je rangovan prave se zbirovi rangova (R_i), dakle, zbirovi rangova po redovima tabele.

-Računamo vrednost izraza $W=12S/[m^2(n^3-n)]$, pri čemu je $S=\sum_{i=1}^{n}(R_i-\frac{\sum_{i=1}^{n}R_i}{n})^2$ suma kvadriranih odstupanja zbirova rangova od očekivane vrednosti sume, m broj kriterijuma rangovanja (ili procenjivača) a n broj entiteta koji se ranguju.[91]
-Kada je veliki broj "vezanih" rangova, tj. kada veliki broj entiteta dobije isti rang bolje je W računati pomoću sledećeg obrasca:

$$W=\frac{S}{(1/12)m^2(n^3-n)-m\sum_{j=1}^{m}T_j}$$

gde je za pojedini kriterijum rangovanja izraz

$$T=\frac{\sum_{v=1}^{p}(t_v^3-t_v)}{12}$$

gde je t broj vezanih rangova koji čine jednu grupu vezanih rangova a p ukupan broj grupa vezanih rangova unutar niza rangova za dati kriterijum rangovanja ili za datog procenjivača.

Testiranje statističke značajnosti W

Ukoliko je n >7 za testiranje nulte hipoteze: $\omega = 0$ (nema slaganja među nizovima rangova) računamo statistik:

$$\chi^2 = m\,(n-1)\,W$$

koji pod H_0 ima Hi-kvadrat distribuciju uzorkovanja sa n-1 stepeni slobode.
Kada je n < 7 radi utvrđivanja značajnosti W koeficijenta treba izračunati egzaktnu verovatnoću statistika ako je H_0 tačna ili konsultovati posebne tablice za to.

Zapamtite: U verziji 8.0 programa SPSS moguće je izračunati egzaktnu verovatnoću pojedinih statistika tako što se pri definisanju procedure, u poddijalogu **Exact tests**, u koji se ulazi klikom na dugme **Exact** u osnovnom dijalogu date procedure, odabere opcija **Exact**.

****

[91] Kada postoji potpuno slaganje rangova koji su dodeljeni prema m kriterijuma za n pojedinih entiteta sume rangova (vrednosti izraza R_i) su m, 2m, 3m,..., nm. Ukupna suma rangova za sve entitete je inače [mn(n+1)]/2. Jasno je da, kada je slaganje maksimalno, variranje suma rangova dodeljenih pojedinim entitetima biva maksimalno i suma kvadriranih zbirova rangova je tada $[m^2(n^3-n)]/12$. Očigledno W je, u suštini, količnik realnog i maksimalnog slaganja nizova rangova dodeljenih svakom entitetu.

Računanje Kendalovog koeficijenta slaganja u programu SPSS vrši se pozivom menija **Statistics/Nonparametric Tests/K Related Samples** i odabirom opcije **Kendall's W** u okviru **Test Type** dijaloga **Tests for Several Related Samples**.
U ispisu se pod naslovom **Kendall Coefficient of Concordance** u koloni **W** daje iznos ovog koeficijenta a u kolonama **Chi Square, D. F.** i **Significance** χ^2 statistik, broj stepeni slobode i verovatnoća greške I tipa pri odbacivanju tačne H_0.
Pri organizaciji podataka za računanje W-koeficijenta treba voditi računa o tome da se entiteti koji se ranguju nalaze u kolonama, a kriterijumi rangovanja (ili procenjivači) u redovima! Ukoliko su podaci uneti na standardni način (entiteti u redovima), pre računanja ovog koeficijenta treba matricu transponovati procedurom **Data/Transpose**.

Primer:
Dvadeset pet ispitanika rangovano je, na osnovu četiri kriterijuma, u pogledu stepena izloženosti traumatskim ratnim iskustvima. Postoji li slaganje između ovih različitih kriterijuma?

```
- - - - - Kendall Coefficient of Concordance

     Mean Rank   Variable
          2.50   I1
          2.75   I2
          5.50   I3
          3.25   I4
          3.25   I5
          6.25   I6
          6.50   I7
          8.50   I8
         10.25   I9
         10.75   I10
         14.25   I11
         13.25   I12
         13.75   I13
         15.50   I14
         15.25   I15
         14.75   I16
         16.25   I17
         20.75   I18
         18.25   I19
         19.25   I20
         17.75   I21
         18.25   I22
         23.25   I23
         24.50   I24
         20.50   I25
          Cases              W        Chi-Square          D.F.   Significance
            4        8.307E-01          79.7446            24       6.68E-08
```

Postoji zavidno slaganje ova četiri kriterijuma u proceni stepena izloženosti ljudi ratnim traumatskim iskustvima. Kendalov koeficijent slaganja (W) je 0.83 i statistički je značajan (p = 0.00000007). U koloni **Mean Rank** dati su prosečni rangovi na osnovu četiri kriterijuma za svakog ispitanika (kolona **Variable**!).
****

XIII.3. Neparametrijske metode za testiranje pretpostavke da uzorci pripadaju populacijama sa istim distribucijama ili populacijama sa istim centralnim vrednostima

Dva nezavisna uzorka

Test medijane

Namena: Njime ispitujemo da li se dva nezavisna uzorka razlikuju u pogledu centralne tendencije, tj. da li (ili ne) pripadaju populaciji sa istom Medijanom.
Uslovi: Podaci moraju poticati bar sa ordinalne merne skale.
Postupak:
-Nalazimo medijanu iz rezultata za oba uzorka zajedno.
-Pravimo tabelu 2x2 gde su kategorije jedne varijable pripadnost jednom ili drugom uzorku, a kategorije druge varijable iznad ili ispod zajedničke medijane.[92]
-Iz ove tabele računamo χ^2 statistik /pod uslovom da je $(n_1 + n_2) > 40$ ili ako je $20 < (n_1 + n_2) < 40$ i sve teorijske frekvence su veće od 5) / ili primenjujemo Fišerov test egzaktne verovatnoće /ako ima teorijskih frekvencija manjih od 5, pri čemu je istovremeno $20 < (n_1 + n_2) < 40$, ili ako je $(n_1 + n_2) < 20$/.

H_0: *Dva uzorka su iz populacije sa istom medijanom.*

Zaključivanje (neodbacivanje ili odbacivanje H_0) je u skladu sa postupkom za Hi-kvadrat test ili Fišerov test egzaktne verovatnoće.

****

Budući da se Test medijane može upotrebiti za testiranje značajnosti razlika dva ili više nezavisnih uzoraka u SPSS programu se izvođenje ovog testa vrši odabirom opcije **Median** u polju **Test Type** dijaloga **Tests for k Independent Samples** menija **Statistics/Nonparametric Tests/K Independent Samples**. U polje **Test Variable List** unosi se ime zavisne varijable a u polje **Grouping Variable** ime kategoričke varijable koja definiše pripadnost subpopulaciji. Za kategoričku varijablu treba popuniti i polje **Define Range**.
U ispisu je pod naslovom **Median Test** data kontingencijska tabela u kojoj je pripadnost poduzorcima (kategorijama kategoričke varijable) ukrštena sa kategorisanom zavisnom varijablom (ispod medijane ili jednako medijani/- **LE median**/ i iznad medijane/- **GT median**/). Ipod tabele su dati ukupan broj slučajeva (**Cases**), zajednička medijana svih rezultata na zavisnoj varijabli (**Median**) i (za velike uzorke) hi-kvadrat statistik (**Chi-Square**) sa verovatnoćom da ovaj statistik bude jednak ili veći dobijenom ako je H_0 tačna (**Significance**) ili (za male uzorke) verovatnoća dobijanja takvog ishoda ako je H_0 tačna iz Fišerovog testa egzaktne verovatnoće (**Exact probability**).

Primer: Ispitanici dveju grupa rangovani su od strane pretpostavljenih u pogledu uspešnosti na poslu. Dve grupe su obučavane različitim metodama za datu vrstu posla. Ispitanici nisu imali nikakvog prethodnog iskustva sa tom vrstom posla.
Da li ove različite metode imaju različite efekte na uspešnost u obavljanju datog posla?

---

[92] Ako se desi da neki rezultat bude jednak medijani možemo ga ili izbaciti (ako je uzorak veliki) ili ga staviti ispod medijane.

- - - - - Median Test

RANG_USP RANG USPEŠNOSTI U OBAVLJANJU POSLA
by GRUPA

		GRUPA 1.000	GRUPA 2.000
RANG_USP	GT median	13	6
	LE median	11	10

Cases	Median	Chi-Square	Significance
40	20.5000	.5054	.4771

Iz tabele se vidi da se ispitanici koji su obučavani poslu različitim metodama skoro podjednako nalaze ispod i iznad zajedničke medijane u pogledu uspešnosti u obavljanju posla. Hi-kvadrat nije statistički značajan pa nemamo razloga da odbacimo pretpostavku o podjednakim efektima obeju primenjenih metoda.

****

Man - Vitnijev (Mann-Whitney) U-test

Namena: Za ispitivanje da li dva uzorka pripadaju populacijama sa istom distribucijom.

Uslovi: Podaci moraju poticati bar sa ordinalne merne skale.
H_0 :Uzorci pripadaju populaciji sa istom distribucijom.

Postupak:

-Rangujemo sve ispitanike zajedno u pogledu ispitivane osobine s tim da očuvamo informaciju o pripadnosti uzorku. Rang jedan dobija algebarski najniži rezultat.

-Za svaki uzorak posebno saberemo rangove i dobijamo sume rangova za svaku grupu ponosob (R_1 i R_2.). Radi kontrole zbir ovih suma (R_1+R_2) mora biti jednak izrazu $[n(n+1)]/2$ gde je n ukupan broj ispitanika.

-Računamo vrednost izraza:

$$U_1 = n_1n_2 + [n_1(n_1+1)/2] - R_1$$

i

$$U_2 = n_1n_2 + [n_2(n_2+1)/2] - R_2$$

-Uzimamo kao statistik U = min (U_1,U_2). Ako je uzorak čije U smo uzeli manji od 20 računamo egzaktnu verovatnoću statistika ili iz tablica očitavamo verovatnoću U ako je tačna H_0. Ako je uzorak čije smo U uzeli kao statistik veći od 20 distribucija uzorkovanja U statistika se približava normalnoj sa očekivanom vrednošću

$$\mu_U = n_1n_2/2$$

i varijansom

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

pa statistik

$$z_{dobijeno} = (U - \mu_U) / \sigma_U$$

ima standardizovanu normalnu distribuciju uzorkovanja.

Ako je $P(|z| \geq z_{dobijeno}) \leq 0.05$ odbacujemo H_0.

****

U SPSS programu se izvođenje ovog testa vrši odabirom opcije **Mann-Whitney U** u polju **Test Type** dijaloga **Two Independent Samples Test** menija **Statistics/Nonparametric Tests/2 Independent Samples**. U polje **Test Variable List** unosi se ime zavisne varijable a u polje **Grouping Variable** ime kategoričke varijable koja definiše pripadnost subpopulaciji. Za kategoričku varijablu treba popuniti i polje **Define Range**.
U ispisu su pod naslovom **Mann-Whitney U - Wilcoxon Rank Sum W Test** dati prosečni rangovi (**Mean Rank**) i broj ispitanika po grupama (**Cases**) a zatim U statistik (**U**) i manja suma rangova (**W**). Ukoliko su uzorci mali data je egzaktna verovatnoća U statistika pod H_0 (**Exact 2-Tailed P**) a za velike uzorke statistik z (**Z**) sa verovatnoćom da ovaj statistik bude jednak ili veći dobijenom ako je H_0 tačna (**2-Tailed P**). Pod imenom **Corrected for ties** dat je statistik z (**Z**) u koji je uključena korekcija za vezane rangove sa verovatnoćom da ovaj statistik bude jednak ili veći dobijenom ako je H_0 tačna (**2-Tailed P**).

Primer: Analiza je urađena na primeru koji je već obrađen Testom Medijane.

```
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test

    RANG_USP  RANG USPEŠNOSTI U OBAVLJANJU POSLA
  by GRUPA


    Mean Rank     Cases

        23.04        24  GRUPA = 1  Metoda 1
        16.69        16  GRUPA = 2  Metoda 2
                     --
                     40  Total

                                Exact              Corrected for ties
          U              W     2-Tailed P            Z       2-Tailed P
        131.0          267.0      .0952          -1.6865        .0917
```

Manja suma rangova u grupi 2 (W) je 267 a U statistik za zaključivanje dobijen je na grupi 1. Bez obzira što ova grupa ima više od 20 ispitanika, s obzirom na to da su uzorci mali, za donošenje zaključka ćemo koristiti egzaktnu verovatnoću. Ta verovatnoća je u ovom slučaju 0.0952 što znači da ne možemo odbaciti H_0. Ni na osnovu ove analize ne možemo odbaciti pretpostavku da dve različite metode obuke ne dovode do razlika u uspešnosti obavljanja datog posla .

****

Više nezavisnih uzoraka

Kraskal-Volisov (Kruskal-Wallis) test.

Namena: Za ispitivanje da li više od 2 nezavisna uzorka pripadaju istoj populaciji u pogledu ispitivane osobine.

Uslovi: Mere moraju poticati bar sa ordinalne merne skale.

H_0 : Uzorci pripadaju istoj populaciji u pogledu ispitivane karakteristike.

Postupak:

-Entiteti svih poduzoraka, tj. kategorija ili nivoa kategoričke varijable ranguju se zajedno tako da najniži rezultat dobije rang 1.
-Za svaki poduzorak ili kategoriju posebno izračunamo sume rangova (R_i). Radi kontrole zbir ovih suma mora biti jednaka izrazu $[n(n+1)]/2$ gde je n ukupan broj ispitanika.
-Računamo statistik $H = [12/n(n+1)] \sum_{i=1}^{k}\left(\frac{R_i^2}{n_i}\right) - 3(n+1)$, gde su R_i sume rangova k poduzoraka, n_i su veličine pojedinih poduzoraka a n ukupan broj rezultata.
-Ako je svaki uzorak veći od 5, statistik H ima Hi-kvadrat distribuciju uzorkovanja sa k-1 stepeni slobode gde je k broj uzoraka.
U tom slučaju računamo $P(\chi^2 \geq H)=\alpha$. Ako je $\alpha < 0.05$ odbacujemo H_0.

Kada su uzorci manji od 5 kritične vrednosti H za odbacivanje H_0 se moraju očitati iz tablica koje se mogu naći u opširnijim udžbenicima neparametrijske statistike.

****

U programu SPSS for Windows se izvođenje ovog testa vrši odabirom opcije **Kruskal-Wallis H** u polju **Test Type** dijaloga **Tests for Several Independent Samples** menija **Statistics/Nonparametric Tests/K Independent Samples**. U polje **Test Variable List** unosi se ime zavisne varijable a u polje **Grouping Variable** ime kategoričke varijable koja definiše pripadnost subpopulaciji. Za kategoričku varijablu treba popuniti i polje **Define Range**.
U ispisu su pod naslovom **Kruskal-Wallis 1-Way Anova** dati prosečni rangovi (**Mean Rank**) i broj ispitanika po grupama (**Cases**) a zatim Hi-kvadrat statistik (**Chi-Square**), broj stepeni slobode (**D.F.**) i verovatnoća greške I tipa pri odbacivanju tačne H_0 (**Significance**). Pod imenom **Corrected for ties** dat je Hi-kvadrat statistik (**Chi-Square**) u koji je uključena korekcija za vezane rangove, broj stepeni slobode (**D.F.**) i verovatnoća greške I tipa pri odbacivanju tačne H_0 **(Significance).**

Primer: Tri grupe disleksičara rangovane su u pogledu izraženosti određene vrste grešaka u čitanju. Rangovanje je izvršeno tako da su svi ispitanici rangovani (bez obzira u kojoj disleksičnoj grupi se nalaze) tako da je ispitanik sa najmanjom izraženošću određene vrste grešaka dobio rang 1. Da li izraženost određene vrste grešaka zavisi od vrste disleksije?

Ispis iz ove procedure izgleda ovako:

```
- - - - - Kruskal-Wallis 1-Way Anova

    RANGBRGR  RANK of BR_GREUC
  by GRUPA     VRSTA DISLEKSIJE


    Mean Rank     Cases

         4.40         5   GRUPA = 1   DUBINSKA
         7.40         5   GRUPA = 2   FONOLOŠKA
        11.50         4   GRUPA = 3   POVRŠINSKA
                     --
                     14   Total

                                                  Corrected for ties
  Chi-Square     D.F.  Significance     Chi-Square        D.F.  Significance
     6.4057       2         .0406         6.4057           2         .0406
```

Hi-kvadrat od 6.41 je u ovom slučaju bio statistički značajan (Significance 0.0406) pa bismo mogli odbaciti nultu hipotezu. Dakle, izraženost date vrste grešaka u čitanju nije ista kod dislleksičara iz ove tri kategorije. Iz kolone Mean Rank možemo da uočimo i da je ova vrsta grešaka najmanje izražena kod tzv. dubinskih disleksičara.

****

Dva zavisna uzorka:

Test predznaka (Sign test)

Namena: Za utvrđivanje da li se dva zavisna uzorka razlikuju u pogledu ispitivane osobine.

H_0: Nema razlika među grupama.

Postupak:

-Svakom paru rezultata dajemo znak + ili - u zavisnosti od toga koji član para je bolji ili lošiji. Ako su članovi para jednaki upisujemo nulu. Pod H_0 očekujemo podjednak broj pluseva i minusa.

-Prebrojimo broj pluseva i broj minusa. Nule se ne uzimaju u obzir!

-Kada je n (broj parova) dovoljno veliko računamo:

$$z = (|D| - 1) / \sqrt{n}$$

gde je D = | broj pluseva- broj minusa | a n broj parova koji imaju ili + ili - predznak. (Nule se ne računaju!). Računamo $P(|z| \geq z_{dobijeno}) = \alpha$. Ako je $\alpha \leq 0.05$ odbacujemo H_0.

-Ako je uzorak mali treba značajnost odrediti uz pomoć tablica, tj. preko binomne raspodele:

$$\sum_{k=n+}^{n} \frac{n!}{k!(n-k)!}\left[\left(\frac{1}{2}\right)^{n-k}\left(\frac{1}{2}\right)^{k}\right]$$

Pomoću gornjeg obrasca moguće je izračunati verovatnoću da se pod H_0 dobije k ili više (do n) pluseva /ili k ili više minuseva/. Ako je dobijena verovatnoća manja ili jednaka 0.05 odbacujemo H_0.

Veliki nedostatak ovog statističkog testa je u tome što on uzima u obzir samo smer razlike, a ne i njenu veličinu. Sem toga, ovaj test se ne može primeniti ako ima veliki broj jednakih parova.

Vilkoksonov (Wilcoxon) test ekvivalentnih parova

Namena: Za utvrđivanje postoji li statistički značajna razlika između dva zavisna uzorka.

Uslovi: Bar intervalna merna skala.

H_0: Uzorci pripadaju istoj populaciji u pogledu ispitivane osobine, tj. nema razlike između grupa.

Postupak:
-Računamo razlike između rezultata koji su u paru i zatim rangujemo razlike (ne uzimajući pri tom u obzir predznak razlike). Pri rangovanju najmanjoj razlici dodeljujemo rang 1. Razlike jednake nuli se ispuštaju!!
-Svakom rangu dodeljujemo predznak koji je imala njemu odgovarajuća razlika.
-Pravimo odvojene sume pozitivnih i negativnih rangova.
-Ako je n> 25 onda R_+ (suma pozitivnih rangova) ima približno normalnu raspodelu sa očekivanom vrednošću $[n(n+1)]/4$ i standardnom devijacijom $\sqrt{[n(n+1)(2n+1)]/24}$ pa se značajnost određuje računanjem statistika z:

$$z = \frac{R_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Statistik z je, dakle, standardizovano (jer je izraz u imeniocu standardna greška za R_+) odstupanje dobijene sume pozitivnih rangova diferencija (R_+) od njene očekivane vrednosti pod nultom hipotezom.
Računamo $P(|z| \geq z_{dobijeno}) = \alpha$. Ako je $\alpha \leq 0.05$ odbacujemo H_0.
Ukoliko je n (broj parova) ≤ 25 u tablicama treba očitati verovatnoću dobijanja takvog ili većeg R_+ pod H_0.

****

Test predznaka i Vilkoksonov test ekvivalentnih parova se u programu SPSS biraju u meniju **Statistics/Nonparametric Tests/K Related Samples** odabirom opcija **Sign** (za test predznaka) i **Wilcoxon** (za test ekvivalentnih parova) u okviru **Test Type** dijaloga **Two-Related Samples Tests**. U dijalogu je potrebno odabrati u levom okviru par rangova (ili varijabli) pa tek potom (klikom na strelicu nadesno) u polje **Test Pair(s) List** ubaciti imena oba člana para.
U ispisu za Test predznaka se pod naslovom **Sign Test** daju brojevi razlika sa negativnim (**- Diffs**) i pozitivnim predznakom (**+ Diffs**),broj razlika jednakih nuli (**Ties**), z-statistik(**Z**) i verovatnoća da z bude veći ili jednak z dobijenom ako je tačna nulta hipotreza (**2-Tailed P**). Za male uzorke data je binomna verovatnoća (**Binomial**).

U ispisu za Test ekvivalentnih parova se pod naslovom **Wilcoxon Matched-Pairs Signed-Ranks Test** daju, prosečni pozitivni i negativni rangovi razlika (**Mean Rank**), brojevi rangova razlika sa negativnim (**- Ranks**) i pozitivnim predznakom (**+ Ranks**),broj razlika jednakih nuli (**Ties**), z-statistik(**Z**) i verovatnoća da z bude veći ili jednak z dobijenom ako je tačna nulta hipotreza (**2-Tailed P**).

Primer: Radi utvrđivanja slaganja novovenčanih bračnih parova u pogledu željene veličine porodice jedan istraživač je zatražio od 12 parova supružnika da odgovore (svaki član bračnog para nezavino od drugog člana para) na pitanje: Koliki je za njih idealni broj dece? Odgovor na pitanje da li se supružnici razlikuju u pogledu ispitivane osobine statistički je potražen primenom Vilkoksonovog testa ekvivalentnih parova i testa predznaka.

```
- - - - - Wilcoxon Matched-Pairs Signed-Ranks Test

     MUZEVI     POŽELJAN BROJ  DECE PO MUŽEVIMA
with SUPRUGE    POŽELJAN BROJ DECE PO SUPRUGAMA

     Mean Rank     Cases

          5.33         6  - Ranks (SUPRUGE LT MUZEVI)
          4.33         3  + Ranks (SUPRUGE GT MUZEVI)
                       3    Ties  (SUPRUGE EQ MUZEVI)
                      --
                      12    Total

         Z =   -1.1255            2-Tailed P =  .2604

- - - - - Sign Test

     MUZEVI     POŽELJAN BROJ  DECE PO MUŽEVIMA
with SUPRUGE    POŽELJAN BROJ DECE PO SUPRUGAMA

           Cases
               6  - Diffs (SUPRUGE LT MUZEVI)
               3  + Diffs (SUPRUGE GT MUZEVI)        (Binomial)
               3    Ties                      2-Tailed P =      .5078
              --
          12  Total
```

Ono što, pre svega, treba uočiti u ispisu je da tri para imaju identično mišljenje (Ties) i da se, stoga, veličina uzorka smanjuje sa 12 na 9! Budući da je n manje od 25 onda verovatnoću da Z statistik u Vilkoksonovom testu bude 1.1255 ili veći od toga (0.2604) ne bismo smeli uzeti sa poverenjem pri odlučivanju da li da odbacimo H_0.
Zato ćemo u ovom primeru odluku doneti na osnovu Testa predznaka. Budući da je verovatnoća da od 9 parova samo na osnovu slučaja bude 6 ili više parova kod kojih supruge žele manje dece no njihovi muževi (ili, što je isto, 3 ili manje parova kod kojih supruge žele više dece no njihovi muževi) jednaka 0.5078 ne bismo mogli da odbacimo nultu hipotezu. Dakle, novopečeni supružnici različitih polova ne razlikuju se u pogledu željenog broja dece.
****

Više zavisnih uzoraka:

Fridmanov (Friedman) test

Namena: Da li više od dve grupe pripadaju istoj populaciji u odnosu na ispitivanu osobinu.
Uslovi: Mere potiču bar sa ordinalne merne skale.

H_0 :Grupe se ne razlikuju u pogledu ispitivane osobine.

Postupak:
-Rezultate svrstavamo u n redova (ispitanici) i k kolona (eksperiment. situacije).
-Rezultati u svakom redu, tj. za svakog ispitanika ranguju se s obzirom na eksperimentalne situacije.
-Saberemo rangove po kolonama, tj. po eksperimentalnim situacijama
/Kontrola rangovanja: $\sum_{j=1}^{k} R_j = \frac{nk(k+1)}{2}$ /gde je R_j zbir rangova za j-tu eksperimentalnu situaciju.
-Računamo statistik:

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

gde su R_j sume rangova po kolonama, n je broj redova a k broj kolona.
Kada je k> 4 (bez obzira na n) ili k=3 i n>9 ili k=4 i n >4 tada χ^2_r ima Hi-kvadrat distribuciju uzorkovanja sa k-1 stepeni slobode. U tim uslovima računamo $P(\chi^2 \geq \chi^2_r) = \alpha$. Ako je $\alpha \leq 0.05$ odbacujemo H_0.

Ako su n i k mali značajnost χ^2_r se očitava iz tablice. Ako je dobijena vrednost χ^2_r jednaka ili veća od vrednosti u tablici odbacujemo H_0.

****

Fridmanov test se u programu SPSS izvodi pozivom menija **Statistics/Nonparametric Tests/K Related Samples** i odabirom opcije **Friedman** u okviru **Test Type** dijaloga **Tests for Several Related Samples**. U ispisu su pod naslovom **Friedman Two-Way Anova** nalaze prosečni rangovi po ekperimentalnim situacijama (**Mean Rank**), broj slučajeva (**Cases**), Hi-kvadrat statistik (**Chi-Square**), broj stepeni slobode (**D.F.**) i verovatnoća greške tipa I pri odbacivanju tačne H_0 (**Significance**).

Primer: Ista grupa od osam ispitanika je učila besmislene slogove pomoću četiri različite tehnike učenja (redosled primene različitih tehnika učenja bio je balansiran). Brojem reprodukovanih slogova merena je uspešnost naučenog. Da li različite tehnike učenja imaju različit efekat na stepen uspešnosti učenja?

- - - - - Friedman Two-Way Anova

Mean Rank	Variable
1.75	TEHN1
2.50	TEHN2
3.63	TEHN3
2.13	TEHN4

Cases	Chi-Square	D.F.	Significance
8	9.4500	3	2.39E-02

Budući da je Hi-kvadrat statistički značajan odbacili bismo nultu hipotezu i zaključili da ove različite tehnike daju različitu uspešnost u učenju date vrste materijala.

****

XIV. MULTIPLA REGRESIONA ANALIZA

XIV.1. Pojam kompozitne (složajne) varijable

Da bi se razumela multipla regresiona analiza kao i multivarijacione statističke metode (diskriminantna, kanonička korelaciona, faktorska analiza...) neophodno je razumeti pojam kompozitne, tj. složajne varijable. Budući da je naglasak u ovom tekstu na linearnim modelima podrobnije ćemo objasniti linearni kompozit, tj. linearnu kombinaciju varijabli.
Kompozitna varijabla sastavljena je od nekoliko sastavnih, komponentnih varijabli. Na primer varijabla U, pri čemu je U =X*Y*Z je **nelinearna** kompozitna varijabla a varijabla $Y = b_1x_1+b_2x_2+\cdots+b_mx_m$ je linearna kompozitna varijabla ili linearna kombinacija varijabli $x_1, x_2,\ldots,x_m$. U izrazu za linearni kompozit $b_1, b_2,\ldots, b_m$ su **koeficijenti** ili **ponderi** za linearnu kombinaciju. Dakle, linearna kombinacija ili linearna kompozitna varijabla nastaje sabiranjem neponderisanih ili ponderisanih komponentnih varijabli. Sami ponderi govore kolika se važnost, tj težina daje pojedinoj varijabli u skupu varijabli koje grade linearnu kombinaciju. Ponderi mogu biti međusobno jednaki ili različiti. Ukoliko su svi ponderi jednaki jedinici tada se kompozit često zove neponderisani kompozit.
Ako su komponentne varijable za linearnu kombinaciju centrirane ili standardizovane, onda se linearna kombinacija varijabli može iskazati u matričnom obliku na sledeći način:

1. Ako je $\mathbf{X}_d$ matrica centriranih ili devijacionih skorova reda nxm i ako je **b** vektor koeficijenata reda mx1, onda se vrednosti ispitanika na linearnoj kombinaciji varijabli iz matrice $\mathbf{X}_d$ nalaze u vektoru **y** reda nx1, pri čemu je $\mathbf{y} = \mathbf{X}_d\mathbf{b}$.

2. Ako je **Z** matrica standardizovanih varijabli reda nxm i ako je $\mathbf{b}^*$ vektor koeficijenata reda mx1, onda se vrednosti ispitanika na linearnoj kombinaciji varijabli iz matrice **Z** nalaze u vektoru $\mathbf{k}^*$ reda nx1, pri čemu je $\mathbf{k}^* = \mathbf{Z}\mathbf{b}^*$.

Aritmetička sredina i varijansa linearnog kompozita

Aritmetička sredina linearne kombinacije varijabli jednaka je zbiru ponderisanih aritmetičkih sredina komponentnih varijabli:

$$M_y = b_1M_1+b_2M_2+\ldots+b_mM_m$$

Ako su varijable $X_1, X_2,\ldots,X_m$ centrirane ($M_1 = M_2 = M_m = 0$), varijansa njihove linearne kombinacije, tj.linearne kompozitne varijable Y je:

$$S_y^2 = \sum_{h=1}^{m}\sum_{j=1}^{m} b_h b_j r_{hj} S_h S_j$$

ili

$$S^2_y = b^2_1S^2_1 + b^2_2S^2_2 +\ldots+ b^2_mS^2_m+2b_1b_2r_{12}S_1S_2 + 2\, b_{m-1}b_mr_{m-1.m}S_{m-1}S_m$$

Pri tom su $b_1, b_2,\ldots b_m$ koeficijenti za ponderisanje sastavnih varijabli, $S^2_1, S^2_2,\ldots,S^2_m$ su njihove varijanse, $S_1, S_2,\ldots, S_m$ standardne devijacije komponentnih varijabli, a r je koeficijent linearne korelacije parova varijabli. Kao što je poznato, izraz $r_{m-1,m}S_{m-1}S_m$ predstavlja kovarijansu m-1 i m-te varijable.

Matrični oblik izraza za varijansu kompozitne varijable, ako su sastavne varijable centrirane, je

$$S^2_y = \mathbf{b}^t\mathbf{C}\mathbf{b}$$

pri čemu je **C** matrica kovarijansi sastavnih varijabli. Ako su, pak, komponentne varijable standardizovane varijansa njihove linearne kombinacije je

$$S^2_k = \mathbf{b}^{*t}\mathbf{R}\mathbf{b}^*$$

gde je **R** matrica interkorelacija sastavnih varijabli.

XIV.2. Multipla regresiona analiza

Struktura podataka

$E = \{ e_i ; i = 1,...,n\} \subset P$ je slučajni uzorak izvučen iz homogene populacije;

$V = \{ v_j ; j = 1,2,..., m\} \subset U$ je skup kvantitativnih prediktorskih varijabli;

u = kvantitativna kriterijumska varijabla.

Uzmimo da smo opisali uzorak *E*, tj. merenjem odredili vrednosti ispitanika u pogledu kriterijumske varijable u, kao i u pogledu skupa prediktorskih varijabli (*V*):

$E \otimes u = \mathbf{y} = (y_i)$
$E \otimes V = \mathbf{X} = (x_{ij})$; i = 1,...,n; j = 1,..., m.

Premda to nije uvek nužno, budući da je multipla regresiona analiza invarijantna na metriku varijabli, napravićemo standardizaciju varijabli tako da su u vektoru **k** (varijabla u) i matrici **Z** (skup prediktorskih varijabli-*V*) podaci u standardizovanom obliku.
Matrica **Z** se iz matrice nestandardizovanih podataka dobija na sledeći način:

$$\mathbf{Z} = (\mathbf{X}-\mathbf{P}\mathbf{X})\,\mathbf{S}^{-1}$$

gde je $\mathbf{P} = \mathbf{e}(\mathbf{e}^t\mathbf{e})^{-1}\mathbf{e}^t$ centroidni projektor (matrica reda n×n čiji su svi elementi 1/n). Proizvod **PX** u izrazu za **Z** je matrica reda n×m u kojoj su kao rezultati ispitanika upisane aritmetičke sredine na datoj varijabli. Budući da je

$$S^2 = \text{diag } \mathbf{C},$$

a

$$\mathbf{C} = (\mathbf{X}^t\mathbf{X} - \mathbf{X}^t\mathbf{P}\mathbf{X})\,(\mathbf{e}^t\mathbf{e})^{-1}$$

je matrica varijansi i kovarijansi, dijagonalna matrica S^{-1} sadrži recipročne vrednosti standardnih devijacija pojedinih varijabli.

Formalna postavka problema multiple regresije i rešenje problema:[93]

$$\mathbf{Zb}^* = \mathbf{k} - \mathbf{g} \mid \mathbf{g}^t\mathbf{g}\, n^{-1} = \min$$

$$\mathbf{b}^* = ?$$

(Oznakom $\mathbf{b}^*$ je u ovom tekstu označen vektor standardizovanih regresionih koeficijenata dok je **b** oznaka vektora nestandardizovanih regresionih koeficijenata).

Izraz **Zb*** može se napisati i u sledećem obliku: $b^*_1 Z_{1i} + b^*_2 Z_{2i} + \ldots + b^*_j Z_{ji}$. U vektoru **g** su greške, tj. razlike između predviđenih (na osnovu linearne kombinacije prediktora) vrednosti kriterijumske varijable (k_i^*) i emirijski dobijenih vrednosti na toj varijabli (k_i). Diferenciranjem funkcije $\mathbf{g}^t\mathbf{g}\, n^{-1} = (\mathbf{k} - \mathbf{Z}\,\mathbf{b}^*)^t(\mathbf{k} - \mathbf{Z}\,\mathbf{b}^*)n^{-1} = (\mathbf{k}^t\mathbf{k}\, n^{-1} - \mathbf{b}^{*t}\mathbf{Z}^t\mathbf{k}\, n^{-1} - \mathbf{k}^t\mathbf{Z}\,\mathbf{b}^* n^{-1} + \mathbf{b}^{*t}\mathbf{Z}^t\mathbf{Z}\,\mathbf{b}^* n^{-1}) = (1 - 2\mathbf{k}^t\mathbf{Z}\,\mathbf{b}^* n^{-1} + \mathbf{b}^{*t}\mathbf{Z}^t\mathbf{Z}\,\mathbf{b}^* n^{-1}) = (1 - 2\mathbf{q}\,\mathbf{b}^* + \mathbf{b}^{*t}\mathbf{R}\,\mathbf{b}^*)$ po elementima vektora $\mathbf{b}^*$ i izjednačavanjem sa nulom dobijamo $\mathbf{q} = \mathbf{R}\,\mathbf{b}^*$, a množenjem cele jednačine sa leva sa $\mathbf{R}^{-1}$ dobijamo rešenje problema:[94]

$$\mathbf{b}^* = \mathbf{R}^{-1}\mathbf{q}$$

gde je $\mathbf{R}^{-1}$ inverz matrice interkorelacija prediktorskih varijabli a **q** vektor korelacija prediktorskih varijabli sa kriterijumskom. Očigledno, matrica **R** mora biti regularna, tj. nesingularna, dakle, $|\mathbf{R}| \neq 0$.

Rečima iskazano, problem je u tome da se nađe takva linearna kombinacija prediktorskih varijabli koja je najbliža (pod kriterijumom najmanjih kvadrata /engl. least square/) empirijski dobijenoj kriterijumskoj varijabli. Dakle, ponderisanje standardizovanih prediktorskih varijabli elementima vektora $\mathbf{b}^*$ daje minimum izraza $\mathbf{g}^t\mathbf{g}\, n^{-1}$. Na taj način postižemo da je $\frac{\sum_{i=1}^{n}(k_i - k_i^*)^2}{n}$ minimalno, tj. prosek kvadriranih odstupanja predviđenih vrednosti kriterijumske varijable (na osnovu takve regresione jednačine) od empirijski dobijenih vrednosti kriterijumske varijable manji je nego što bi to bio slučaj sa bilo kakvim drugačijim ponderima.[95]

Na osnovu dobijenog rešenja možemo izraziti i kvadrat koeficijenta multiple korelacije (koeficijent multiple determinacije):

$$R^2_{y.1,2..m} = \mathbf{b}^{*t}\mathbf{R}\,\mathbf{b}^* = \mathbf{q}^t\mathbf{R}^{-1}\mathbf{q}$$

Ako su prediktorske i kriterijumska varijabla stnadardizovane, onda je koeficijent multiple determinacije jednak varijansi predviđene kriterijumske varijable:

$$S^2_{k^*} = k^{*t}k^*(1/n) = \mathbf{b}^{*t}\mathbf{Z}^t\mathbf{Z}\,\mathbf{b}^*(1/n) = \mathbf{b}^{*t}\mathbf{R}\,\mathbf{b}^* = \mathbf{q}^t\mathbf{R}^{-1}\mathbf{q} = R^2_{y.1,2,..m}$$

---

[93] Prikaz multiplog regresionog modela bez korišćenja matrične notacije može se pročitati u Vuković (1997), str. 563 - 618.

[94] Budući da su i kriterijumska i prediktorske varijable standardizovane jasno je da je matrica interkorelacija prediktorskih varijabli **R** jednaka $\mathbf{Z}^t\mathbf{Z}n^{-1}$, vektor korelacija prediktorskih varijabli sa kriterijumskom (označen kao **q**) jednak je $\mathbf{Z}^t\mathbf{k}\, n^{-1}$ ili $\mathbf{k}^t\mathbf{Z}n^{-1}$, a varijansa kriterijumske varijable ($\mathbf{k}^t\mathbf{k}\, n^{-1}$) jednaka jedinici.

[95] Ukoliko se radi o nestandardizovanim varijablama vektor pondera (**b**) dobija se na sledeći način:
$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$

Komponente ukupnog varijabliteta kriterijumske varijable (sume kvadriranih odstupanja)

Ukupni varijabilitet kriterijumske varijable (definisan sumom kvadriranih odstupanja pojedinačnih rezultata od aritmetičke sredine na kriterijumskoj varijabli y i označen sa SS_t) možemo razbiti na dve komponente:

$$\sum_{i=1}^{n}(y_i - M_y)^2 = \sum_{i=1}^{n}(y_i^* - M_y)^2 + \sum_{i=1}^{n}(y_i^* - y_i)^2$$

ili

$$SS_t \quad = \quad SS_r \quad + \quad SS_g$$

gde je:

SS_t - Suma kvadriranih odstupanja rezultata od AS svih rezultata na kriterijumskoj varijabli, tj. ukupna suma kvadriranih odstupanja;

SS_r - Suma kvadriranih odstupanja za regresiju, tj. suma kvadriranih odstupanja predviđenih (na osnovu regresionog modela) vrednosti kriterijumske varijable od AS rezultata na kriterijumskoj varijabli. Naravno, aritmetička sredina dobijenih i predviđenih vrednosti kriterijumske varijable je ista.

SS_g - Suma kvadriranih odstupanja predviđenih od empirijski dobijenih rezultata na kriterijumskoj varijabli, suma kvadrata grešaka predviđanja (ili reziduala).

Koeficijent multiple korelacije (i determinacije) iskazani preko suma kvadriranih odstupanja

Koeficijent multiple determinacije je količnik sume kvadriranih odstupanja za regresiju kriterijumske varijable na skup prediktorskih varijabli (deo varijabiliteta kriterijumske varijable koji je predviđen, tj.objašnjen regresijom) i ukupne sume kvadriranih odstupanja kriterijumske varijable (ukupni varijabilitet kriterijumske varijable):

$$R^2_{y.1,2,..,m} = \frac{SS_r}{SS_t} = \frac{\sum_{i=1}^{n}(y_i^* - M_y)^2}{\sum_{i=1}^{n}(y_i - M_y)^2}$$

gde je M_y aritmetička sredina kriterijumske varijable.
Koeficijent multiple korelacije pokazije kolika je povezanost između kriterijumske varijable, sa jedne, i linearne kombinacije prediktorskih varijabli (predviđena kriterijumska varijabla), sa druge strane, a kvadrat multiple korelacije govori o tome koliki procenat varijanse kriterijumske varijable možemo objasniti (predvideti) na osnovu poznavanja varijanse prediktorskih varijabli.

Koeficijent multiple determinacije (kvadrat koeficijenta multiple korelacije) dobijen na uzorku obično precenjuje populacijsku vrednost i to utoliko više ukoliko ima više

prediktorskih varijabli a manje ispitanika. Stoga je poželjno, da bi se dobila nepristrasna ocena, izvršiti korekciju ovog statistika kada on služi kao ocena populacijskog koeficijenta determinacije :

$$\text{Korigovani } R^2_{y.1,2,...,m} = 1 - \frac{n-1}{n-m-1}(1 - R^2_{y.1,2,...,m})$$

Koeficijent multiple korelacije iskazan preko koeficijenata korelacije prediktora i kriterijumske varijable i standardizovanih regresionih koeficijenata:

$$R_{y.1.2...m} = \sqrt{b_1^* r_{yx1} + b_2^* r_{yx2} + \ldots + b_m^* r_{yxm}}$$

ili

$$R^2_{y.1.2...m} = \mathbf{b}^{*t}\mathbf{q}$$

a budući da je $\mathbf{b}^* = \mathbf{R}^{-1}\mathbf{q}$ onda je $R^2_{y.1.2...m} = \mathbf{q}^t\mathbf{R}^{-1}\mathbf{q}$.

Koeficijent multiple determinacije iskazan preko nestandardizovanih regresionih koeficijenata i kovarijansi prediktorskih i kriterijumske varijable:

$R^2_{y.1.2...m} = \mathbf{b}^t\mathbf{u}S^{-2}_y = \mathbf{u}^t\mathbf{C}^{-1}\mathbf{u}S^{-2}_y$, pri čemu su u vektoru **u** kovarijanse pojedinih prediktorskih varijabli i kriterijumske varijable, **C** je matrica kovarijansi prediktora, a S^2 je varijansa kriterijumske varijable.

Koeficijent multiple korelacije je zapravo Pirson-Braveov koeficijent korelacije između empirijski dobijene kriterijumske varijable i predviđene kriterijumske varijable. Koeficijent multiple determinacije se može posmatrati i kao proporcionalno smanjenje greške predviđanja kriterijumske varijable na osnovu linearne kombinacije prediktora u odnosu na grešku predviđanja kada ne posedujemo informacije o prediktorskim varijablama.

Standardna greška predviđanja (ili ocene)

Standardna greška predviđanja je standardna devijacija distribucije reziduala, tj. mera prilagođenosti regresionog modela podacima :

$$SE_{y.1.2...m} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-m-1}} = S_y\sqrt{1 - R^2_{y.1,2,...,m}}$$

Kao i svaka standardna devijacija, standardna greška je iskazana u istim jedinicama u kojima su dati i rezultati na kriterijumskoj varijabli i govori o tome koliko je dati regresioni model podesan kao model datih podataka, tj. koliko dobro model odgovara (engl. fit) podacima. Standardna devijacija kriterijumske varijable je pokazatelj prosečne greške koju pravimo kada (pod kriterijumom najmanjih kvadrata) predviđamo rezultate na kriterijumskoj varijabli nemajući nikakve informacije o rezulatima ispitanika na prediktorskim varijablama. Budući da je aritmetička sredina varijable tačka za koju važi da je suma kvadriranih odstupanja od te tačke minimum, tj. najmanja moguća, najbolje (sa stanovišta principa najmanjih kvadrata) predviđanje u tom slučaju je da za svakog ispitanika predviđeni rezultat bude aritmetička sredina kriterijumske varijable. Zato

standardnu grešku predviđanja treba posmatrati u odnosu prema standardnoj devijaciji kriterijumske varijable. Što je standardna greška predviđanja manja (u odnosu na standardnu devijaciju kriterijumske varijable) to će prosečna greška u predviđanju rezultata na kriterijumskoj varijabli na osnovu regresionog modela, tj. poznavanja rezultatata na prediktorskim varijablama, biti manja u odnosu na prosečnu grešku koju bismo napravili u predviđanju bez poznavanja rezultata na prediktorima.

Testiranje hipoteze o populacijskom koeficijentu multiple korelacije:

$H_0 : P_{y.1,2,..m} = 0$

Nulta hipoteza tvrdi da je koeficijent multiple korelacije u populaciji jednak nuli tj. da nema linearne povezanosti kriterijumske varijable, sa jedne, i prediktorskih varijabli uzetih zajedno, sa druge strane.

Test statistik

$$f = \frac{R^2_{y.1,2,..m} / m}{1 - R^2_{y.1,2,..m} / n - m - 1}$$

ima, ako je H_0 tačna, Snidikorovu F-distribuciju uzorkovanja sa m i n-m-1 stepeni slobode. ($R^2_{y\ 1,2,..m} = \mathbf{b}^* {}^t\mathbf{R}\ \mathbf{b}^* = \mathbf{q}^t\mathbf{R}^{-1}\mathbf{q}$ je kvadrat koeficijenta multiple korelacije, tj. koeficijent multiple determinacije, dobijen na uzorku, m je broj prediktorskih varijabli a n je veličina uzorka).
Ukoliko ne odbacimo H_0 sledi da su sve korelacije prediktorskih varijabli sa kriterijumskom u populaciji jednake nuli. Ako smo, pak, odbacili H_0 moguće je testirati značajnost koeficijenata linearne korelacije između pojedinog prediktora i kriterijuma:

$H_0: \rho_{y.j} = 0$

Test statistik $f_j = r^2_{y.xj}\ (n-2)/\ 1 - r^2_{y.xj}$ ima Snidikorovu F- distribuciju uzorkovanja sa 1 i n-2 stepeni slobode.
Ovaj statistički test podrazumeva da kriterijumska varijabla ima normalnu a prediktorske varijable multivarijatnu normalnu funkciju gustine.

Pretpostavke regresionog modela koje stoje u osnovi testiranja statističke značajnosti koeficijenta multiple korelacije prikazanim F-testom:

1. a) nezavisnost reziduala, tj. nezavisno izvlačenje jedinica uzorka;
   b) normalnost raspodele reziduala;
   c) jednakost varijansi (homoscedastičnost) reziduala za sve kombinacije nezavisnih varijabli.
2. a) Normalnost raspodele kriterijumske varijable;
   b) Homogenost varijansi (homoscedastičnost) kriterijumske varijable za sve kombinacije prediktorskih varijabli.

Regresioni koeficijenti

Pokazuju parcijalni uticaj pojedinih prediktorskih varijabli na kriterijumsku varijablu kada se svi ostali prediktori drže konstantnim. Njihova veličina govori o tome za koliko se menja kriterijumska varijabla ako se dati prediktor promeni za jednu jedinicu, a ostali prediktori ostanu nepromenjeni. Ukoliko su nestandardizovani (vektor **b**), regresioni koeficijenti se mogu direktno međusobno porediti samo ako su skale prediktorskih varijabli iste[96].
Odnos nestandardizovanih i standardizovanih regresionih koeficijenata:

$$b_j = b^*_j * S_y / S_{xj}$$

gde su S_y i S_x standardne devijacije kriterijumske i prediktorske varijable.
Standardizovani regresioni koeficijenti proporcionalni su parcijalnim korelacijama uniknih (anti-image) komponenti prediktora i kriterijuma[97]. Očigledno je da kada postoji samo jedna prediktorska varijabla standardizovani regresioni koeficijent je jednak koeficijentu linearne korelacije kriterijumske i prediktorske varijable.

Testiranje hipoteze o populacijskim regresionim koeficijentima:

H_0 (j): $\beta_j = 0$; j = 1, 2,..., m

Test statistik $t_j = b_j / SE_{bj}$ (gde je SE_{bj} standardna greška regresionog koeficijenta) ima Strudentovu t-distribuciju uzorkovanja sa n-2 stepeni slobode.
Ishode statističkih testova hipoteza o regresionim koeficijentima treba uzeti obazrivo jer oni nisu međusobno nezavisni budući da je matrica kovarijansi regresionih koeficijenata

$$\mathbf{C}_{b^*} = (1 - R^2_{y.1,2,...,m})\, \mathbf{R}^{-1}\, [1/(n-m-1)]$$

nedijagonalna[98].
Da bi se izbegli veliki titraji i dobile stabilne ocene regresionih koeficijenata potrebno je da razlika između veličine uzorka i broja prediktorskih varijabli bude velika (bar 200).
Regresioni faktor:

Regresioni faktor je vektor koji sadrži korelacije između pojedinih prediktora i predviđene kriterijumske varijable. Ovaj vektor je istovremeno jednak vektoru korelacija prediktorskih varijabli i kriterijumske varijable čiji je svaki element podeljen koeficijentom multiple korelacije:

$$\mathbf{f}_r = \mathbf{Z}^t \mathbf{k}\, \mathbf{n}^{-1} (R_{y.1,2...m})^{-1} = \mathbf{q}\, (R_{y.1,2...m})^{-1}$$

---

[96] Za poredjenje regresionih koeficijenata u slučaju kada ovaj uslov nije ispunjen pojedini autori predlažu parcijalni koeficijent elastičnosti:

$$l_j = b_j M_j / M_y.$$

Ovaj koeficijent pokazuje za koliko procenata se menja kriterijumska varijabla kada se vrednost neke prediktorske varijable promeni za 1 % a ostali prediktori ostaju nepromenjeni.

[97] O pojmu parcijalne korelacije može se videti u odeljku **XIV. 2.** ovog teksta.

[98] Matrica kovarijansi nestandardizovanih regresionih koeficijenata je
$\mathbf{C}_b = S^2_g\, (\mathbf{X}^t\mathbf{X})^{-1}$
pri čemu je **X** matrica nestandardizovanih podataka na prediktorskim varijablama, a S^2_g, ocena varijanse reziduala, tj. varijanse greške.

Zapamtite:

- Koeficijent multiple korelacije ne može biti niži od apsolutne vrednosti najveće korelacije između pojedinog prediktora i kriterijumske varijable.
- Svaka dodatna prediktorska varijabla nužno dovodi do toga da koeficijent multiple korelacije ostane isti ili se poveća, u zavisnosti od toga da li dodatni prediktor sadrži varijansu koja nije zajednička sa ostalim prediktorima u modelu.
- Da bi se na osnovu uzoračke vrednosti dobila što nepristrasnija ocena koeficijenta multiple korelacije u populaciji potrebno je da veličina uzorka ispitanika daleko premašuje (bar 20 do 30 puta) broj prediktorskih varijabli.To je neophodno jer je očekivana vrednost koeficijenta multiple determinacije, kada je koeficijent multiple korelacije u populaciji jednak nuli: $E(R^2_{y.1,2,..m}) = (m-1) / (n-1)$, pri čemu je m broj prediktora a n veličina uzorka. Očigledno je da sa približavanjem broja prediktora veličini uzorka, očekivana vrednost koeficijenta multiple determinacije teži jedinici, iako je populacijski koeficijent korelacije jednak NULI!
- Ukoliko se, pri analizi regresionog modela, uopšte postavi problem važnosti doprinosa pojedinih prediktorskih varijabli objašnjenju, tj.predviđanju varijabiliteta kriterijumske varijable, možda najadekvatnije rešenje za taj problem je da se posmatra veza koju sažima koeficijent multiple korelacije kao veza između kriterijumske varijable i skupa prediktora kao celine. Teorijski je često korisnije ne dekomponovati tu vezu i dodeljivati "bodove" izvornim prediktorskim varijablama već pokušati sa tumačenjem sadržaja nove kompozitne varijable preko tumačenja sadržaja regresionog faktora[99].

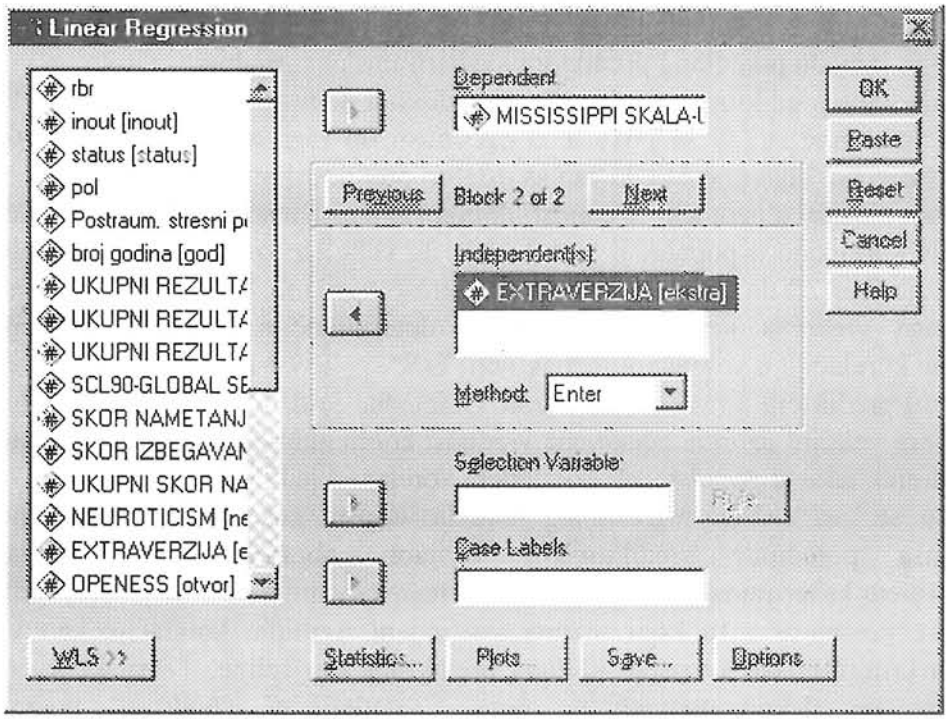
****

U programu SPSS for Windows potrebno je odabrati meni **Statistics/Regression/Linear** i u dijalog prozoru **Linear Regression** u okvir **Dependent** ubaciti ime kriterijumske a u okvir **Independent(s)** imena prediktorskih varijabli. Polje **Method** ostaje kako je po default-u, tj. **Enter** (ubacivanje svih prediktorskih varijabli u model u jednom koraku). U poddijalogu **Statistics** može se odabrati željeni statistik. Od statistika treba odabrati bar **Descriptives** (daje AS i SD varijabli i matricu interkorelacija varijabli), **Estimate** u okviru **Regression Coefficient** (ocene regresionih koeficijenata) i **Model Fit** (testiranje značajnosti koeficijenta multiple korelacije). U opciji **Save** određujemo šta treba sačuvati u aktivnom fajlu (poželjno je, mada nije uvek nužno, sačuvati nestandardizovane ili standardizovane predviđene vrednosti **/Unstandardized** ili **Standardized** u okviru **Predicted values/** i nestandardizovane ili standardizovane reziduale/**Unstandardized** ili **Standardized** u okviru **Reziduals/**). U **Options** je po default-u definisano da se ispitanici sa nedostajućim vrednostima na samo jednoj varijabli izbace potpuno iz analize (**Exclude cases listwise**). Ukoliko se odabere opcija **Replace with mean** onda će nedostajuće vrednosti ispitanika na nekoj varijabli biti zamenjene aritmetičkom sredinom date varijable.
Veoma korisna mogućnost u definisanju ove procedure je postupno uvođenje prediktora u model. To se postiže tako da se definiše model sa jednim ili više prediktora na prethodno objašnjen način a zatim klikne na dugme **Next** iznad okvira **Independent(s)** u osnovnom prozoru (slika 8). Situacija prikazana na slici 8 je dobijena je pošto je prethodno definisan regresioni model sa rezultatom na Mississippi skali kao merom kriterijumske varijable i dva prediktora (na prethodnom prozoru njihova imena su bila ubačena u okvir **Independent(s)** , a zatim je klikom na dugme **Next** u model ubačen i treći prediktor (Ekstraverzija).

---

[99] Da bi se razumeo pojam dodatnog doprinosa svakog sledećeg prediktora u modelu potrebno je razumevanje pojma semiparcijalne korelacije o kojem se može pročitati u odeljku **XIV.1.** ovog teksta.

Slika 8: Prozor za dijalog za definisanje multiple regresije u SPSS 8.0



U verziji 8.0 programa moguće je zatražiti još parcijalnu i semiparcijalnu korelaciju (klikom na dugme **Statistics** i uključivanjem opcije **Part and Partial Correlations** u dijalogu **Linear Rgression: Statistics**).

Na početku ispisa dati su (posle AS, SD i matrice interkorelacija varijabli) koeficijent multiple korelacije (**Multiple R**), kvadrat ovog koeficijenta (**R Square**), korigovani koeficijent multiple determinacije (nepristrasna ocena populacijske vrednosti -**Adjusted R Square**) i standardna greška (**Standard error**) a potom u bloku **Analysis of Variance** test značajnosti koeficijenta multiple determinacije. Suma kvadrata za regresiju (**Regression** u koloni **Sum of Squares**), i suma kvadrata reziduala (**Residual**) podeljeni njima pripadajućim brojem stepeni slobode (za regresiju df=m a za reziduale df=n-m-1) daju prosečne kvadrate odstupanja (**Mean square**). Količnik prosečnog kvadrata za regresiju i prosečnog kvadrata za reziduale (grešku) je **F**-statistik. Vrednost **Signif F** je verovatnoća da se na slučajnom reprezentativnom uzorku date veličine dobije toliki F-statistik ili još veći ako je nulta hipoteza tačna. U bloku **Variables in the Equation** je data tabela sa nestandardizovanim (**B**) i standardizovanim regresionim koeficijentima (**BETA**). Količnik B/SE B daje **T** statistik, a u koloni **sig T** su rezultati testiranja značajnosti pojedinih regresionih koeficijenata. **Constant** je intercept (odsečak), tj. vrednost kriterijumske varijable kada su vrednosti svih prediktorskih varijabli jednake nuli. Radi dobijanja regresionog faktora potrebno je sačuvati vrednosti kriterijumske varijable predviđene na osnovu regresione jednačine a potom u meniju **Statistics/Correlate/Bivariate** napraviti korelacije prediktorskih varijabli i predviđene kriterijumske varijable.

Budući da je jednofaktorska univarijatna analiza varijanse (i analiza varijanse uopšte) samo specijalni slučaj multiple regresione analize (tj. kanoničke korelacione analize) u kojoj je kriterijumska varijabla kategoričkog tipa, u SPSS programu se jednofaktorska analiza varijanse može izvesti i pozivanjem menija **Statistics/ Regression/Linear**. Pre toga je potrebno kategoričku, tj. nezavisnu varijablu kodirati u dummy oblik, tj. pretvoriti je u k-1

binarnih varijabli (gde je k broj kategorija).[100] U dijalogu **Linear Regression** se kao prediktorske varijable (polje **Independent(s)**) unosi k-1 dummy (dami) ili binarnih varijabli a kao kriterijumska (polje **Dependent**) kvantitativna, zavisna varijabla iz analize varijanse. Konstanta (**Intercept**) će u tom slučaju biti jednaka aritmetičkoj sredini one kategorije (grupe) koja na svim dummy, tj. binarnim varijablama ima vrednost nula. Dodavanjem nestandardizovanog regresionog koeficijenta na konstantu dobija se aritmetička sredina kategorije koja na dummy varijabli čiji je to koeficijent ima vrednost 1. Testovi značajnosti regresionih koeficijenata (ne uzimajući u obzir test za intercept) pokazuju značajnost razlika ostalih grupa u odnosu na onu grupu čija aritmetička sredina je jednaka interceptu. Nekorigovani koeficijent multiple korelacije jednak je u tom slučaju η-koeficijentu a test značajnosti koeficijenta multiple korelacije identičan je F-testu iz procedure **One-Way ANOVA**.

Primer:
Na uzorku izbeglica u našoj zemlji prikupljeni su, upitnikom za procenu ratnih stresora, podaci o ukupnoj izloženosti ratnim stresorima (borbenim, van linije fronta, zarobljeničkim i sl.). Ukupna relativna izloženost ratnim stresorima (variabla E) iskazana je kao proporcija u odnosu na broj od 50 reprezentativnih ratnih stresora. Osim toga, Mississipi skalom izmerena je izraženost PTSD simptoma. Upitnikom NEOPI-R izmerene su i bazične crte ličnosti. S obzirom na to da smo na ovom primeru (u odeljku o jednostavnoj regresiji) odgovorili na pitanje u kojoj meri se individualne razlike u izraženosti simptoma PTSD-a u izbegličkoj populaciji mogu predvideti ili objasniti linearnom vezom izrazenosti PTSD simptomatologije sa ekstenzitetom izloženosti ratnim stresorima, ovde ćemo pokušati da ustanovimo koliko uspešno možemo da predvidimo razlike u postraumatskim reakcijama (izražene razlikama u izraženosti PTSD simptomatologije) kada u skup prediktora osim ekstenziteta izloženosti stresoru uključimo i dve bazične crte ličnosti: neuroticizam i ekstraverziju.

Ispis iz verzije 6.1. programa SPSS (sva tri prediktora su ubačena u model):

```
* * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data

              Mean   Std Dev   Label

MSUKUP       96.767   19.190  MISSISSIPPI SKALA-UKUPAN SKOR
E              .191     .151  RELATIVNA UKUPNA IZLOŽENOST STRESORIMA
NEUROT      153.552   19.140  NEUROTICISM
EKSTRA      134.569   18.505  EXTRAVERZIJA

N of Cases =    116
```

[100] Ovo kodiranje se može obaviti iz menija **Transform/Recode/Into Different Variable**. Za svaku grupu, osim jedne, definiše se varijabla na kojoj će ispitanici koji spadaju u datu kategoriju imati vrednost 1 a svi ostali 0. Ispitanici iz grupe koja nema odgovarajuću dummy varijablu imaće na svim dummy varijablama vrednost 0. Radi se zapravo o pretvaranju kategoričke varijable u selektorsku matricu koja sadrži jednu kolonu manje od broja kategorija na kategoričkoj varijabli.

Correlation, 1-tailed Sig:

	MSUKUP	E	NEUROT	EKSTRA
MSUKUP	1.000	.321	.532	-.319
	.	.000	.000	.000
E	.321	1.000	-.049	.064
	.000	.	.299	.246
NEUROT	.532	-.049	1.000	-.406
	.000	.299	.	.000
EKSTRA	-.319	.064	-.406	1.000
	.000	.246	.000	.

* * * * M U L T I P L E R E G R E S S I O N * * * *

Listwise Deletion of Missing Data

Equation Number 1 Dependent Variable.. MSUKUP MISSISSIPPI SKALA-UKUPAN SK

Block Number 1. Method: Enter E NEUROT EKSTRA

Variable(s) Entered on Step Number
1.. EKSTRA EXTRAVERZIJA
2.. E RELATIVNA UKUPNA IZLOŽENOST STRESORIMA
3.. NEUROT NEUROTICISM

Multiple R	.64834
R Square	.42034
Adjusted R Square	.40481
Standard Error	14.80465

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	3	17800.82962	5933.60987
Residual	112	24547.88590	219.17755

F = 27.07216 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
E	45.163214	9.189684	.354411	4.915	.0000
NEUROT	.492848	.078955	.491578	6.242	.0000
EKSTRA	-.147155	.081736	-.141904	-1.800	.0745
(Constant)	32.287902	19.492348		1.656	.1004

End Block Number 1 All requested variables entered.

* *

Residuals Statistics:

	Min	Max	Mean	Std Dev	N
*PRED	70.3021	142.4114	96.7672	12.4415	116
*RESID	-31.1153	39.8424	.0000	14.6103	116
*ZPRED	-2.1272	3.6687	.0000	1.0000	116
*ZRESID	-2.1017	2.6912	.0000	.9869	116

Total Cases = 208

* *

From Equation 1: 2 new variables have been created.

Name	Contents
PRE_1	Predicted Value
ZRE_1	Standardized Residual

Ispis iz verzije 8.0 programa SPSS (sa sva tri prediktora u modelu):

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	EKSTRA, E, NEUROT	,	Enter

a All requested variables entered.
b Dependent Variable: MSUKUP

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,648	,420	,405	14,8046

a Predictors: (Constant), EKSTRA, E, NEUROT

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17800,830	3	5933,610	27,072	,000
	Residual	24547,886	112	219,178		
	Total	42348,716	115			

a Predictors: (Constant), EKSTRA, E, NEUROT

b Dependent Variable: MSUKUP

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Correlations Zero-order	Partial	Part
1	Constant	32,288	19,492		1,656	,100			
	E	45,163	9,190	,354	4,915	,000	,321	,421	,354
	NEUROT	,493	,079	,492	6,242	,000	,532	,508	,449
	EKSTRA	-,147	,082	-,142	-1,800	,074	-,319	-,168	-,130

a Dependent Variable: MSUKUP

Koeficijent multiple korelacije od 0.65 je statistički značajan. Dakle, između skupa prediktora uzetih zajedno, sa jedne, i izraženosti simptomatologije PTSD-a sa druge strane postoji linearna povezanost u populaciji. Od ukupnog varijabiliteta u pogledu ove kriterijumske varijable 40% varijabiliteta (**Adjusted R Square** je .40481) možemo objasniti na osnovu toga što se izbeglice razlikuju u pogledu osobina ličnosti i stepena izloženosti ratnim stresorima. Uočimo da smo uvođenjem bazičnih crta ličnosti u skup prediktora postigli povećanje objašnjenog varijabiliteta na kriterijumskoj varijabli sa blizu 10%, koliko smo u stanju da predvidimo na osnovu modela koji uključuje samo stepen izloženosti ratnim stresorima, na 40%! Iz matrice interkorelacija svih varijabli koja se u ispisu nalazi pod naslovom **Correlation, 1-tailed Sig:** vidimo da osobine ličnosti koreliraju sa kriterijumskom varijablom ali ne i sa stepenom izloženosti ratnim stresorima (što bismo i očekivali).

U ispisu iz verzije 8.0 programa, model za predviđanje rezultata ispitanika na Mississippi skali, tj. za predviđanje izraženosti simptoma PTSD-a bio bi:

Predviđeni MS skor = 32.29 + 45.16*E + 0.49*NEUROT -0.15*EKSTRA.

Ukoliko pogledamo parcijalne standardizovane regresione koeficijente (kolona BETA) možemo zaključiti da najveći parcijalni doprinos objašnjenju individualnih razlika u pogledu izraženosti PTSD simptomatologije ima osobina neuroticizma. U kontekstu ostalih prediktora parcijalni doprinos ekstraverzije nije statistički značajan (t statistik je 1.80 a verovatnoća greške prvog tipa pri odbacivanju pogrešne nulte hipoteze da je parcijalni regresioni koeficijent u populaciji jednak nuli je 0.0745), što znači da bismo mogli ovaj prediktor i isključiti iz skupa prediktorskih varijabli a da to ne umanji bitno valjanost modela. Ukoliko to uradimo dobićemo sledeći ispis:

```
* * * *  M U L T I P L E  R E G R E S S I O N  * * * *

Listwise Deletion of Missing Data
Equation Number 1     Dependent Variable..    MSUKUP    MISSISSIPPI SKALA-
UKUPAN SK

Block Number  1.  Method:  Enter       E         NEUROT
Variable(s) Entered on Step Number
   1..     NEUROT     NEUROTICISM
   2..     E          RELATIVNA UKUPNA IZLOŽENOST STRESORIMA
Multiple R             .63527
R Square               .40356
Adjusted R Square      .39301
Standard Error       14.95075
```

```
Analysis of Variance
                    DF       Sum of Squares      Mean Square
Regression           2          17090.39734       8545.19867
Residual           113          25258.31817        223.52494

F =      38.22929        Signif F =  .0000
------------------ Variables in the Equation ------------------
Variable                B         SE B        Beta          T  Sig T

E              44.359685     9.269424     .348105      4.786  .0000
NEUROT           .550315      .072928     .548897      7.546  .0000
(Constant)      3.814379    11.506629                   .331  .7409

End Block Number   1   All requested variables entered.
```

Očigledno, model bez uzimanja u obzir ekstraverzije, skoro podjednako je dobar kao i model sa ovim dodatnim prediktorom. Ukoliko od kvadrata multiple korelacije, koji je dobijen sa ekstraverzijom kao trećim prediktorom (pored neuroticizma i ukupne izloženosti stresorima) i koji iznosi 0.42034 oduzmemo kvadrat multiple korelacije koji se dobija bez tog prediktora u modelu i koji je jednak 0.40356 dobićemo kvadrat koeficijenta semiparcijalne korelacije (pogledati odeljak **XIV.3.**) izloženosti ratnim stresorima i ekstraverzije. Kvadrat koeficijenta semiparcijalne korelacije u ovom slučaju je 0.01678, što znači da uvođenjem ekstraverzije u skup prediktora procenat objašnjene varijanse u kriterijumskoj varijabli povećavamo samo za nešto manje od 2% (ovo dobijamo kao 0.016778*100).

Da je to zaista tako, možemo se uveriti i na osnovu ispisa iz procedure u kojoj je ekstraverzija naknadno ubačena u model kao treći prediktor pored ukupne izloženosti stresorima i neuroticizma (v. objašnjenje slike 8).

Ispis iz ovako definisane procedure (SPSS 8.0) izgleda ovako:

```
U ovom ispisu je Model 1 onaj u kojem su prediktori izloženost stresorima
i neuroticizam a Model 2 je, pored ova dva, kao prediktor uključivao i
ekstraverziju.
Variables Entered/Removed

                    Model    Variables Variables  Method
                               Entered   Removed
                        1    NEUROT, E         ,  Enter
                        2       EKSTRA         ,  Enter

a  All requested variables entered.

b  Dependent Variable: MSUKUP

Model Summary

            R       R Adjusted R      Std.        Change Statistics
               Square     Square  Error of
                                      the
                                 Estimate
   Mode                                   R Square       F df1  df2 Sig. F
                                            Change Change          Change
        ,635    ,404       ,393   14,9508    ,404 38,229   2  113   ,000
        ,648    ,420       ,405   14,8046    ,017  3,241   1  112   ,074

a  Predictors: (Constant), NEUROT, E
```

b Predictors: (Constant), NEUROT, E, EKSTRA

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17090,397	2	8545,199	38,229	,000
	Residual	25258,318	113	223,525		
	Total	42348,716	115			
2	Regression	17800,830	3	5933,610	27,072	,000
	Residual	24547,886	112	219,178		
	Total	42348,716	115			

a Predictors: (Constant), NEUROT, E
b Predictors: (Constant), NEUROT, E, EKSTRA
c Dependent Variable: MSUKUP

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig	Correlations Zero-order	Partia	Part
1	Constant	3,814	11,507		,331	,74			
	E	44,360	9,269	,348	4,786	,00	,321	,41	,348
	NEUROT	,550	,073	,549	7,546	,00	,532	,57	,548
2	Constant	32,288	19,492		1,656	,10			
	E	45,163	9,190	,354	4,915	,00	,321	,42	,354
	NEUROT	,493	,079	,492	6,242	,00	,532	,50	,449
	EKSTRA	-,147	,082	-,142	-1,800	,07	-,319	-,16	-,130

a Dependent Variable: MSUKUP

Excluded Variables

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	EKSTRA	-,142	-1,800	,074	-,168	,833

a Predictors in the Model: (Constant), NEUROT, E
b Dependent Variable: MSUKUP

Ključni deo prethodnog ispisa je kolona **Change Statistics** u tabeli **Model Summary**. Promena koeficijenta multiple determinacije (**R square change**) od 0.017 govori nam da je samo oko 2% dodatne varijanse objašnjeno u modelu 2, tj. uvođenjem ekstraverzije u prediktore pored onih koji su ubačeni u modelu 1 (ukupna relativna izloženost stresorima i neuroticizam). Ne samo da je ova promena koeficijenta determinacije mala već nije ni statistički značajna (F za promenu - **F change** je 3.241, a značajnost ovog F je u koloni **Sig. F change** i iznosi 0.074). To je i razlog zašto je automatski program izbacio ekstraverziju iz modela (tabela **Excluded Variables**). U koloni **Part** tabele **Coefficients** je koeficijent semiparcijalne korelacije između kriterijumske varijable i ekstraverzije i on iznosi -0.13. Kvadrat ovog koeficijenta jednak je promeni koeficijenta multiple determinacije do koje dolazi pri uvođenju ekstraverzije u skup prediktora.

Budući da smo, u analizi sa sva tri prediktora, sačuvali vrednosti na predviđenoj kriterijumskoj varijabli pod imenom **PRE_1**, u meniju **Statistics /Correlate/ Bivariate** možemo da napravimo regresioni faktor, tj. vektor koeficijenata korelacija između pojednih prediktora i predviđene kriterijumske varijable:

```
                    - -  Correlation Coefficients  - -

              E           NEUROT      EKSTRA

PRE_1          .5135        .8553       -.4982
              ( 194)      ( 194)      ( 194)
              P= .000     P= .000     P= .000
 Coefficient / (Cases) / 2-tailed Significance)
```

Na osnovu regresionog faktora možemo zaključiti da veliki stepen izloženosti ratnim stresorima nosi rizik visoke izraženosti PTSD simptoma, ako je udružen, pre svega, sa visokom neuroticizmom i (što je manje važno) izraženom introverzijom.

****

XIV.3. Koeficijenti parcijalne i semiparcijalne korelacije

Pri analizi multiplog regresionog modela često je korisno poznavati "čistu" korelaciju između kriterijumske varijable i pojednog prediktora kada se svi ostali prediktori drže konstantnim. Isto tako, potrebno je, pri dodavanju novog prediktora u model analizirati dodatni doprinos novog prediktora objašnjenju varijabiliteta kriterijumske varijable. U te svrhe se mogu upotrebiti parcijalna i semiparcijalna korelacija.

Parcijalna korelacija je postupak računanja 'čiste' korelacije između dveju varijabli kada se isključi uticaj treće varijable (koja je s njima povezana) ili nekoliko ostalih varijabli koje su sa njima povezane, tj kada se ostale varijable (jedna ili više) statistički drže konstantnim. Ovim postupkom se, zapravo, često mogu otkriti tzv.lažne ili besmislene korelacije.

Kada je reč samo o tri varijable, ako su poznati koeficijenti linearne korelacije među varijablama x, y i z, koeficijent parcijalne korelacije između varijabli x i y se računa prema sledećem obrascu:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

Za računanje matrice parcijalnih korelacija, kada su "u igri" više od tri varijable potrebno je poznavati matricu antiimaž (antiimage) varijabli.

Antiimaž prostor[101]

Ako je matrica **E** takva da je:

$$\mathbf{E} = \mathbf{Z}\mathbf{R}^{-1}\mathbf{U}^2 \text{ gde je } \mathbf{U}^2 = (diag\ \mathbf{R}^{-1})^{-1}$$

ta matrica je dijagonalna matrica koja predstavlja greške predviđanja svake varijable na osnovu m -1 preostalih varijabli (matrica uniknih varijansi). Na taj način, preslikavamo

[101] Podrobnije objašnjenje o imažu i antiimažu varijable može se pročitati u Momirović, Wolf i Popović (1999), str. 65-69.

podatke iz matrice podataka u novu matricu tako da umesto varijabli kakve imamo pretvaramo svaku varijablu u onaj njen deo (antiimaž varijable) koji nema ničeg zajedničkog sa preostalim varijablama. Svaka varijabla ima svoj antiimage inače ne bi postojala, mogla bi se izraziti preko drugih varijabli.

Imaž varijable

Imaž varijable je onaj njen deo koji ona deli sa preostalim varijablama. Matrica varijabli u imaž obliku je:

$$\mathbf{G} = \mathbf{Z} - \mathbf{E} = \mathbf{Z}\,(\mathbf{I} - \mathbf{R}^{-1}\mathbf{U}^2)$$

Kod merenja, **E** kazuje koliku grešku pravimo jer je:

$$\mathbf{E}^t\mathbf{E} * 1/n = \mathbf{U}^2\,\mathbf{R}^{-1}\mathbf{U}^2$$

dok je:

$$\mathbf{G}^t\mathbf{G} * 1/n = \mathbf{R} + \mathbf{U}^2\,\mathbf{R}^{-1}\mathbf{U}^2 - 2\,\mathbf{U}^2$$

Matrica parcijalnih korelacija sadrži antiimaž varijable sa obrnutim predznakom:

$$\mathbf{R}_{parc} = \mathbf{U}^2 - \mathbf{U}^2\,\mathbf{R}^{-1}\mathbf{U}^2$$

Značajnost parcijalne korelacije

H_0: $\rho_{xy.z} = 0$ (Koeficijent parcijalne korelacije u populaciji je jednak 0).

H_1: $\rho_{xy.z} \neq 0$

H_0 se testira pomoću t-statistika sa N-3 stepeni slobode:

$$t = \frac{r_{xy.z}}{\sqrt{(1 - r^2_{xy.z}) / n - 3}}$$

****

Koeficijent parcijalne korelacije dveju varijabli (x i y) ako se isključi uticaj treće varijable (z) može se napraviti u meniju **Statistics/Correlate/Partial** tako što se u prozoru za dijalog **Partial Correlations** u okvir **Variables** unesu imena x i y varijable a u okvir **Controlling for:** ime varijable z. U ispisu iz ove procedure dati su (kao elementi simetrične matrice) koeficijenti parcijalne korelacije, broj stepeni slobode (u zagradi) i verovatnoća greške prvog tipa pri odbacivanju nulte hipoteze ako je parcijalna korelacija u populaciji jednaka nuli (**p=**). Pored ove mogućnosti, u novijim verzijama programa (8.0) parcijalna korelacija se može dobiti unutar procedure multiple regresije (klikom na dugme **Statistics** i uključivanjem opcije **Part and Partial Correlations** u dijalogu **Linear Regression: Statistics**). U ispisu iz regresione analize koeficijent parcijalne korelacije nalazi se u koloni **Partial** tabele **Coefficients**.

Ukoliko napravimo parcijalnu korelaciju između neuroticizma (varijabla NEUROT) i izraženosti PTSD simptomatologije (varijabla MSUKUP), uz statističku kontrolu ukupne

izloženosti ratnim stresorima (varijabla E) i ekstraverzije, na podacima na kojima smo radili multiplu regresionu analizu dobijamo sledeći rezultat:

```
- P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S -

Controlling for..    EKSTRA    E

                MSUKUP       NEUROT

MSUKUP          1.0000         .5080
                (    0)      ( 112)
                P= .         P= .000


NEUROT            .5080       1.0000
                ( 112)       (    0)
                P= .000      P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Parcijalna korelacija je 0.51 i skoro je identična Pirsonovom koeficijentu korelacije između ovih dveju varijabli koji je, kao što smo ranije videli 0.53. Ovo ne čudi, s obzirom na to da neuroticizam praktično ne korelira ni sa jednom od varijabli koje smo držali konstantnim (varijable E i EKSTRA).

****

Semiparcijalna korelacija

Korelacija x varijable i reziduala y varijable koji preostaje kada se iz y varijable izbaci deo varijanse koji ona deli sa trećom varijablom. Iz poznatih korelacija između varijabli x, y i z semiparcijalna korelacija između varijabli x i y računa se na sledeći način:

$$r_{x(y\,z)} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{yz}^2)}}$$

Kvadrat semiparcijalne korelacije govori o tome koliko dodatne varijanse kriterijumske varijable možemo da objasnimo ukoliko varijablu x uvedemo kao dodatni prediktor u regresioni model. Tako, ukoliko je x varijabla uvedena kao prediktor u model u kojem su već dva prediktora, kvadrat semiparcijalne korelacije varijable x i kriterijumske varijable je:

$$R^2_{y.1,2,x} - R^2_{y.1,2}$$

F-testom je moguće testirati da li je ovaj dodatni doprinos varijable x kao trećeg prediktora (pored prediktora 1 i 2) statistički značajan.

F statistik:

$$F = [(R^2_{y.1.2.x} - R^2_{y.1.2})/(m_1 - m_2)]/[(1 - R^2_{y.1.2.x})/(n - m_1 - 1)]$$

(gde je n veličina uzorka, m_1 broj prediktora u modelu sa x varijablom a m_2 je broj prediktora bez x varijable) ima pod H_0 Snidikorovu F-distribuciju uzorkovanja sa $df_1 = (m_1 - m_2)$ i $df_2 = (n - m_1 - 1)$.

****

Kvadrat semiparcijalne korelacije može se u SPSS programu dobiti u meniju **Statistic/Regression/Linear** kao sastavni deo multiple regresione analize. Ukoliko želimo da izračunamo kvadrat semiparcijalne korelacije između varijable X_2 i Y i da na taj način ustanovimo koliki procenat **dodatne** varijanse Y varijable se može objasniti uvođenjem varijable X_2 u regresioni model u kojem je već kao prediktor uvedena varijabla X_1, u starijim verzijama programa potrebno je napraviti dve regresione analize: u prvoj bi X_1 bila prediktor, a u drugoj analizi pored X_1 uvodimo istovremeno X_2 kao **drugu** prediktorsku varijablu. Razlika koeficijenta multiple determinacije iz druge i prve analize je kvadrat semiparcijalne korelacije između X_2 i Y.

U verziji 8.0 programa koeficijent semiparcijalne korelacije moguće je zatražiti u proceduri za definisanje regresione analize klikom na dugme **Statistics** i uključivanjem opcije **Part and Partial Correlations** u dijalogu **Linear Regression: Statistics**. U ispisu iz regresione analize koeficijent parcijalne korelacije nalazi se u koloni **Part**tabele **Coefficients**.

****

XV. ANALIZA GRUPISANJA

U početnim fazama istraživanja određene oblasti često se pokušava sa nalaženjem tipologija entiteta koji se proučavaju. Tako su, na primer, u psihologiji vršeni pokušaji tipologije ljudi prema fizičkim (građa tela) ili psihološkim (osobine temperamenta, crte ličnosti i sl.) karakteristikama. Analiza grupisanja je egzaktan način da se dođe do tiplogije ili da se postojeće tipologije provere.
Cilj analize grupisanja, koja se još zove i taksonomska ili klasterska analiza, je razvrstavanje, tj.klasifikacija entiteta (ili varijabli) u što srodnije (prema određenim obeležjima) grupe, taksone ili klastere. Dakle, cilj nam je da nađemo grupe koje su takve da:

1. Budu što više međusobno udaljene u višedimenzionalnom prostoru;
2. Entiteti unutar iste grupe budu međusobno što sličniji.

Očigledno je da ova dva kriterijuma mogu često da budu sukobljena, a u taksonomskoj analizi nastojimo da pomirivši ih nađemo optimalno rešenje.

(U ovom tekstu će biti objašnjena samo analiza grupisanja sa kvantitativnim varijablama, premda je taksonomska analiza moguća i na osnovu kategoričkih, tj. nekvantitativnih varijabli)[102].

Struktura podataka:

$E = \{e_i,\ i = 1,\ldots,n\} \subset P = \bigcup_k^g P_k$ je slučajni uzorak izvučen iz populacije koja čini uniju g subpopulacija koje po pretpostavci prirodno postoje.

$V = \{\ v_j\ ;\ j = 1,2,\ldots,\ m\} \subset U$ je skup kvantitativnih varijabli na osnovu kojih će se vršiti grupisanje.
Matrica **X** je matrica podataka dobijenih merenjem ispitamika iz skupa E u pogledu varijabli iz skupa V:

$E \otimes V = \mathbf{X} = (x_{ij});\ i = 1,\ldots,n;\ j = 1,\ldots,\ m.$

$Q = \{\ q_k\ ;\ k = 1,2,\ldots,g\}$ je kategorička varijabla koja ima g kategorija i koja definiše pretpostavljene subpopulacije. Za varijablu Q samo pretpostavljamo da postoji i da ćemo je dobiti kao ishod grupisanja.

Formalna postavka problema:

Formalno posmatrano, potrebno je na osnovu matrice **X**, u kojoj su rezultati entiteta na skupu varijabli V da generišemo matricu **S**:
$\mathbf{S} = (s_{ik});\ i = 1,\ldots,n;\ k = 1,\ldots,\ g$. Matrica **S** je selektorska matrica dobijena tako da je:

$$s_{ik} = 1 \mid e_i \in q_k$$
$$s_{ik} = 0 \mid e_i \notin q_k$$

[102] O metodama grupisanja koje koriste binarne ili nominalne podatke može se pročitati u Ferligoj, A. (1989).

Matrica S je matrica koja definiše pripadnost entiteta jednoj od pretpostavljenih subpopulacija, tj. jednoj od kategorija kategoričke varijable Q. U matrici S entiteti koji su svrstani u istu grupu trebalo bi da budu međusobno što sličniji u pogledu skupa varijabli V a da, istovremeno, grupe međusobno budu što različitije.
Ovaj problem spada u tzv. optimizacione probleme nad skupom dozvoljenih grupisanja: Ako je R* neko grupisanje (razvrstavanje) treba napraviti grupisanje R* tako da bude $P(R^*) = \min_{C \in \Phi} P(R)$, pri čemu je P neka kriterijumska funkcija, C je dato grupisanje, a Φ skup dopustivih grupisanja. Kriterijumska funkcija je određena izmerenim distancama među entitetima.
Ovako postavljen problem, mada teorijski rešiv, praktično najčešće nije moguće egzaktno rešiti. Ako je broj entiteta veoma mali (manji od 15), može se, pregledom svih mogućih dopustivih grupisanja, odabrati ono koje je najviše u skladu sa datom kriterijumskom funkcijom. Praktična teškoća ovako postavljenog problema je u tome što sa povećanjem broja entiteta broj mogućih grupisanja enormno raste. Nije moguće, naime, izvršiti sva moguća grupisanja pa potom odabrati, u skladu sa postavljenim kriterijumom, ono koje je najbolje. Broj svih mogućih grupisanja n entiteta u g grupa jednak je Stirlingovom broju:

$$S(n, g) = \frac{1}{g!}\sum_{i=0}^{k}(-1)^{g-i}\binom{g}{i} i^n$$

Stoga se ovaj problem najčešće rešava heuristički, tj. dolazi se do približnog rešenja koje ne mora biti i najbolje.

Dve ključne stvari koje treba rešiti pre samog izvođenja grupisanja su:
1. Odabir mere distance ili sličnosti među entitetima
2. Odabir metode grupisanja

Ad1.

Mera distance
Sve metode grupisanja započinju računanjem matrice bliskosti entiteta. Ukoliko je matrica bliskosti matrica distanci u glavnoj dijagonali ove (kvadratne i simetrične) matrice reda n x n (gde je n broj entiteta koji se grupišu) su nule.
Da bi neka mera bila smatrana rastojanjem (d) među entitetima e_r i e_s potrebno je da ispunjava sledeće uslove:
a) $d(e_r, e_s) \geq 0$ (nenegativnost);
b) $d(e_r, e_r) = 0$. Rastojanje objekta od samog sebe je nula.
c) $d(e_r, e_s) = 0 \mid e_r = e_s$
d) $d(e_r, e_s) = d(e_s, e_r)$ (simetričnost). Rastojanje od e_r do e_s jednako je rastojanju od e_s do e_r.
e) $(\forall t)\ d(e_r, e_s) \leq d(e_r, e_t) + d(e_t, e_s)$ (nejednakost trougla)
Postoji veliki broj poznatih mera rastojanja. Najpoznatija rastojanja za intervalne podatke su:

a) Rastojanje Minkowskog:

$$d_{rs}^{f} = \left(\sum_{j=1}^{m}\left|x_{rj} - x_{sj}\right|^{f}\right)^{1/f} \text{ za } f \neq 0$$

b) Kvadrirano Euklidsko rastojanje :

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^t (\mathbf{x}_r - \mathbf{x}_s)$$

gde su $\mathbf{x}_r$ i $\mathbf{x}_s$vektori podataka za entitete e_r i e_s na m varijabli.
c) Mahalanobisova udaljenost:

$$M_{rs} = (\mathbf{x}_r - \mathbf{x}_s)^t \mathbf{C}^{-1} (\mathbf{x}_r - \mathbf{x}_s)$$

gde su $\mathbf{x}_r$ i $\mathbf{x}_s$vektori podataka za entitete e_r i e_s na m varijabli a $\mathbf{C}^{-1}$ inverz matrice kovarijansi m varijabli. Dakle, ovo rastojanje uzima u obzir i eliminiše iz računanja rastojanja povezanost među varijablama. Kada je matrica korelacija među varijablama matrica identiteta Mahalanobisova udaljenost jednaka je kvadriranom Euklidovskom rastojanju.
Specijalni slučajevi rastojanja Minkovskog su:
-Euklidsko rastojanje za koje je f =2;
-Menhetn rastojanje za koje je f =1;
-Čejbiševljeva distanca za koju je f = ∞

Mera sličnosti

Da bi $s(e_r, e_s)$ bila mera sličnosti entiteta e_r i e_s treba da ispunjava sledeće uslove:

a) $0 \leq s(e_r, e_s) \leq 1$ za $\forall r$ i $\forall s$ (normiranost);
b) $s(e_r, e_s) = s(e_s, e_r)$ (simetričnost);
c) $s(e_r, e_s) = 0 \mid e_r = e_s$. Ako su e_r i e_s identični objekti onda je sličnost među njima jednaku jedinici.

Najpoznatije mere sličnosti za intervalne podatke su:

a) Kosinusni koeficijent $s_{rs} = \dfrac{\sum_{j=1}^{m} x_{rj} x_{sj}}{\sqrt{\sum_{j=1}^{m} x_{rj}^2} \sqrt{\sum_{j=1}^{m} x_{sj}^2}}$. To je kosinus ugla između vektora rezultata na varijablama iz V za entitete e_r i e_s . Izraz u brojiocu je suma unakrsnih produkata elemenata ovih vektora, a koreni u imeniocu su dužine vektora.
b) Pirsonov produkt-moment koeficijent ili koeficijent linearne korelacije.

Da bi se ovi koeficijenti koristili kao mera sličnosti (budući da se kreću u rasponu od -1 do +1) uzima se njihova apsolutna vrednost ili se na njihov iznos dodaje jedinica pa se dobijeni zbir podeli sa 2.

Ad b) Metode grupisanja
Dele se na hijerarhijske, nehijerarhijske i metode zasnovane na ekstremizovanju parsimonijske funkcije.

Hijerarhijske metode

Hijerarhijske metode mogu biti aglomerativne ili metode udruživanja (entiteti se grupišu u sve manji broj klastera da bi se na kraju svi entiteti svrstali u jednu grupu) i metode deobe ili divizivne (klaster koji obuhvata sve entitete se deli na sve veći broj klastera sve dok svaki entitet ne postane klaster za sebe). Pri svakom koraku u ovim metodama, dobijamo različito klastersko rešenje, u pogledu broja i sastava klastera. Najveći problem je odabrati

optimalno, tj. "pravo" rešenje. Pri tome, izbor jednog rešenja kao optimalnog treba da bude vođen teorijom i principima parsimonije (ekonomičnosti) i interpretabilnosti. U svakom novom koraku, tokom aglomerativnog grupisanja, traži se najmanja od sledeće tri vrste distanci:

1. distanca dva najsličnija klastera;
2. distanca između postojećih klastera i negrupisanih entiteta;
3. distanca dva najsličnija negrupisana entiteta.

Sledeći korak, dakle, zavisi od toga koja od ove tri distance je najmanja.

Od mnogih hijerarhijskih metoda objasnićemo samo one koje se nalaze u programu SPSS for Windows. Razlike među metodama su u načinu na koji određuju distance između klastera ili između klastera i nesvrstanog entiteta.

-Metoda **međugrupnog povezivanja (Between-groups linkage)** pri računanju distanci dva klastera uzima u obzir sve distance među parovima entiteta (svaki član para je iz različitog klastera). Tako, ako su entiteti 1 i 2 i 3 iz klastera A, a entiteti 4 i 5 iz klastera B, distanca između klastera A i B se određuje računanjem proseka distanci između parova (1, 4) (1, 5) (2, 4) (2, 5) (3, 4) (3, 5). Dakle, distanca između dva klastera dobija se kao prosek distanci između svih mogućih parova entiteta, pri čemu su članovi para iz **različitih** klastera. Pri udruživanju klastera nastoji se da se maksimizuje distanca među svim parovima entiteta iz dva **različita** klastera.

-U metodi **unutargrupnog povezivanja (Within-groups linkage)** distanca među klasterima jednaka je prosečnoj distanci između svih mogućih parova entiteta koji se dobijaju kada se klasteri spoje u isti klaster. Dakle, distanca klastera A i B ako se ova dva klastera udružuju u jedan računa se kao prosečna distanca između svih parova entiteta koji će činiti potencijalni novi klaster. Pri udruživanju se nastoji da klasteri unutar sebe budu što homogeniji, tj. da prosek distanci parova entiteta unutar novonastalog klastera bude što manji.

-U metodi **najbližeg suseda (Nearest neighbor)** ili jednostrukog povezivanja (Single linkage) disanca među klasterima odgovara distanci dva najbliža entiteta iz dva različita klastera.

-Metoda **najdaljeg suseda (Furthest neighbor)** ili potpunog povezivanja (Complete linkage) distancu među grupama izražava preko najveće distance između dva entiteta iz različith grupa.

-U **centroidnoj metodi (Centroid clustering)** ili Metodi ponderisanih centroida (Weighted pair-group centroid) mera međusobnog rastojanja dva klastera je distanca između centroida, tj.između aritmetičkih sredina za svih m varijabli ova dva klastera. Pri tome je centroid novodobijenog klastera ponderisana (veličinama klastera) kombinacija centroida dva spojena klastera.

-Metoda medijane (Median clustering) ili metoda neponderisanih centroida (Unweighted pair-group centroid) je ista kao metoda centroida sem što je centroid klastera dobijen spajanjem dve grupe neponderisana kombinacija centroida udruženih klastera. Na taj način, klasteri rezličite veličine daju jednak doprinos centroidu.

-U **Vordovoj metodi grupisanja (Ward's method)** prvo se računaju aritmetičke sredine svakog klastera na svakoj od m varijabli, a zatim suma kvadriranih Euklidovih distanci svakog entiteta od aritmetičkih sredina grupe kojoj entitet pripada. Pri spajanju klastera nastoji se da ukupna suma ovih distanci (tj. suma kvadrata unutar svih grupa) bude što manja.

Zapamtite:
Metode **centroida, medijane** i **Vordova** metoda zahtevaju korišćenje **kvadriranih Euklidovih distanci** kao mera rastojanja među entitetima.

Lens i Vilijems (Lance & Wiliams) su pokazali da se većina hijerarhijskih metoda razvrstavanja mogu tretirati kao specijalni slučajevi metode u kojoj se udaljenja između klastera definišu na sledeći način:

$$d(C_i \cup C_j, C_k) = \alpha_1 d(C_i, C_k) + \alpha_2 d(C_j, C_k) + \beta d(C_i, C_j) + \gamma | d(C_i, C_k) - d(C_j, C_k) |$$

gde su C_i, C_j i C_k klasteri, d je distanca klastera a α_1, α_2, β i γ su koeficijenti čijim izborom možemo dobiti određenu metodu. Na primer, kada su $\alpha_1 = 1/2$, $\alpha_2 = 1/2$, $\beta = 0$ i $\gamma = -1/2$ reč je o metodi najbližeg suseda.

Određivanje broja klastera pri hijerarhijskom grupisanju

Određivanje optimalnog broja taksona pri hijerarhijskom grupisanju je jedan od najvećih problema u primeni ovih metoda. Postoji mnoštvo formalnih kriterijuma koji se primenjuju u ove svrhe[103].

a) Najpoznatiji, i jedan od kriterijuma sa najboljim osobinama, je indeks koji su definisali Calinski & Harabasz:

$$[\text{trag } \mathbf{B}/(k-1)]/[\text{trag } \mathbf{W}/(n-k)]$$

gde n predstavlja ukupan broj entiteta, k je broj kalstera, **B** je matrica suma kvadrata i unakrsnih proizvoda između klastera a **W** matrica suma kvadrata i unakrsnih proizvoda unutar klastera. Optimalni broj klastera odgovara onom nivou hijerarhije za koji je C & H indeks maksimalan;

b) C-index:

$$C = [d_w / \min(d_w)]/[\max(d_w) - \min(d_w)].$$

Oznakom d_w označena je suma distanci između entiteta unutar klastera. Optimalni broj klastera je onaj za koji je C-index najmanji.

[103] Podroban prikaz ovih i mnogih drugih kriterijuma može se naći u Milligan, G.W. & Cooper, M.C. (1985).

c) Adaptirani Gudman-Kraskalov gama koeficijent:

$$g = [s(+) - s(-)] / [s(+) + s(-)]$$

gde je s(+) je broj konzistentnih poređenja, tj. onih u kojima su unutarklasterske distance parova entiteta veće od međuklasterskih, a s(-) je broj nekonzistentnih poređenja. Maksimalna vrednost ovog koeficijenta na određenom nivou, tj. za rešenje sa određenim brojem klastera, pokazuje nivo u hijerarhijskom grupisanju na kojem se treba zadržati.

d) F-količnik za testiranje hipoteze o postojanju c_1 prema c_2 klastera, pri čemu je $c_1 < c_2$. F-testom /sa $df_1 = p*(c_1 - c_2)$ i $df_2 = m*(n-c_2)$ stepeni slobode/ poredimo porast u proseku kvadriranih odstupanja od centroida klastera kada se broj klastera smanji sa c_2 na c_1 klastera ($MS_{w\,(c2\,-c1)}$) u odnosu na prosek kvadriranih odstupanja kada je broj klastera c_2 ($MS_{w\,(c2\,)}$):

$$F = MS_{w\,(c2\,-c1)} / MS_{w\,(c2\,)}$$

Nažalost, nijedan od prikazanih kriterijuma nije ugrađen u procedure za grupisanje koje postoje u programu SPSS.

Nehijerarhijske metode

K-means metoda ili metoda k-proseka

U ovoj metodi se počinje sa k (slučajnih) klastera ili se unapred definiše nekoliko tačaka u m-dimenzionom prostoru koje mogu odgovarati i stvarnim entitetima. To su inicijalni centri (centroidi) klastera. Pri tome k, tj.broj klastera definiše korisnik unapred. Potom se entiteti pridružuju klasteru čijem centroidu su najbliži, tj. u odnosu na čiji centroid imaju najmanju euklidsku distancu. Zatim se u svakom koraku računaju centroidi klastera i vrši, ako je neophodno, premeštanje entiteta iz klastera u klaster do stabilizovanja procesa, tj.dok se ne postigne optimalno rešenje, dok se ne minimizuje varijablilitet unutar klastera a , istovremeno, maksimizuje varijablitet između klastera. U ovom procesu kao da se nastoji da se premeštanjem entiteta iz klaster u klaster postigne što veći F količnik iz analize varijanse u kojoj je klasterska pripadnost faktor, tj. nezavisna varijabla. Jedan od kriterijuma valjanog razvrstavanja je stepen u kojem je na osnovu finalnih udaljenosti centroida klastera moguće reprodukovati inicijalna udaljenja između entiteta.

Na kraju grupisanja moguće je proveriti objektivnost dobijenog razvrstavanja (grupisanjem drugih slučajnih uzoraka iz iste populacije) i stabilnost rešenja (ustanovljavanjem da li posle uvođenja novih entiteta taksoni ostaju otprilike isti) a isto tako prirodu razlika među grupama diskriminantnom analizom sa taksonskom pripadnošću kao kategoričkom varijablom.

****

U programu SPSS se grupisanje hijerarhijskim metodama izvodi odabirom menija **Statistics/Classify/Hierarchical Cluster** a potom se u dijalogu **Hierarchical Cluster Analysis** kao **Variables** unesu imena varijabli po kojima se grupišu entiteti a u polje **Label Cases by** može se (a ne mora) uneti ime **tekstualne** varijable koja (ako postoji u fajlu sa podacima) definiše identifikaciju entiteta na ispisu. U okviru **Cluster** odabira se **Cases** ako

se želi grupisanje entiteta ili **Variables** ukoliko se želi grupisanje varijabli. U okviru **Display** određuje se da li na ispisu očekujemo samo statistike (uključena opcija **Statistics**) ili i crteže (uključena opcija **Plots**). Pritiskom na opciju **Statistics** dobija se dijalog **Hierarchical Cluster Analysis: Statistics** u kojem možemo (zavisno od potreba) odabrati da se u ispisu prikaže proces grupisanja iz koraka u korak (**Agglomeration Schedule**), matrica distanci (**Distance matrix**) i pripadnost entiteta klasteru (bilo samo za jedno rešenje grupisanja - polje **Single Solution** ili za više različitih rešenja -polje **Range of Solutions**). Ukoliko se klikne na opciju **Single Solution** u prazno polje treba upisati broj koji definiše koliko grupa će imati rešenje za koje želimo da nam se na ispisu prikaže klasterska pripadnost entiteta. Ako želimo da se na ispisu pojavi informacija o klasterskoj pripadnosti entiteta za sva rešenja od 2 grupe pa do rešenja sa 6 grupa u prvo prazno polje (posle **From**) opcije **Range of Solutions** upisujemo broj 2 a u drugo (posle reči **through**) broj 6.

Ako želimo da vidimo i grafički ispis iz ove procedure (uključena opcija **Plots** u okviru **Display**) u dijalogu **Hierarchical Cluster Analysis: Plots** možemo odabrati hijerarhijsko drvo ili dendrogram (**Dendrogram**) ili **Icicle** za sva moguća grupisanja (**All Clusters** u okviru **Icicle**) ili samo za rešenja sa određenim brojem grupa (**Specified Range of Clusters**). Ukoliko ne želimo crtež **Icicle** treba odabrati opciju **None** u tom okviru. U okviru **Orientation** možemo za crtež **Icicle** definisati da li želimo njegovu vertikalnu ili horizontalnu orijentaciju.

Odabir mere rastojanja i metode grupisanja vrši se u dijalogu **Hierarchical Cluster Analysis: Method** (u ovaj dijalog ulazimo klikom na polje **Method** iz osnovnog dijaloga). U dijalogu **Method** biramo željenu metodu tako što kliknemo na strelicu nadole iza polja **Cluster Method** čime otvaramo spisak mogućih metoda. Odabir metode koju želimo vršimo prostim klikom na ime metode u spisku.

U okviru **Measure** po istom principu biramo željenu meru rastojanja.[104] Jedina razlika je u tome što ovde još pre izbora same mere moramo specifikovati da li je reč o meri rastojanja za intervalne podatke (**Interval**), frekvencije (**Counts**) ili binarne (**Binary**) podatke. Ukoliko su podaci u varijablama na osnovu kojih vršimo grupisanje intervalnog ili frekvencijskog tipa možemo ih transformisati odabirom neke od mogućnosti u okviru **Transform Values**. Mere rastojanja se takođe mogu transformisati u željeni oblik odabirom jedne od opcija u okviru **Transform Measures**.

Samo ukoliko je reč o grupisanju entiteta (a ne varijabli) moguće je u dijalogu **Hierarchical Cluster Analysis: Save** definisati da se sačuva klasterska pripadnost entiteta kao nova varijabla u aktivnom fajlu i to bilo za jedno (opcija **Single Solution** u okviru **Cluster Membership**) ili za više rešenja (opcija **Range of Solutions**).

Ispis iz procedure **Hierarchical Cluster Analysis** sadrži sledeće elemente:

-Posle informacije o broju entiteta i vrsti odabrane mere rastojanja data je (ukoliko je tražena) matrica distanci (**Dissimilarity Coefficient Matrix**) a potom pod naslovom **H I E R A R C H I C A L C L U S T E R A N A L Y S I S** prikazan je, korak po korak, proces grupisanja (**Agglomeration Schedule using **** Method**). U koloni **Coefficient** su mere distanci u datom koraku.

---

[104] Osim mera rastojanja objašnjenih u tekstu, program SPSS omogućuje i da sami definišemo distancu između dva entiteta za intervalne podatke kao r-ti koren iz sume apsolutnih razlika vrednosti entiteta na varijablama , podignutim na p-ti stepen. U okviru **Measure** dijaloga **Hierarchical Cluster Analysis: Method** odaberemo opciju **Interval** a zatim u polju s desne strane odaberemo **Customized** a u ulazna polja **Power** (stepen) i **Root** (koren) upišemo željene vrednosti. Pri tom stepen i koren mogu uzeti vrednosti 1 ili 2 ili 3 ili 4. Očigledno je da, ako odaberemo stepen 2 i koren 2. dobijamo Euklidovu distancu.

-Ukoliko je to zahtevano pri definisanju procedure, u nastavku ispisa date su informacije o klasterskoj pripadnosti svakog entiteta pod naslovom **Cluster Membership of Cases using *** Method** a potom i crteži (prvo **Icicle Plot using *** Method** pa onda **Dendrogram**) koji grafički pokazuju proces grupisanja.
-Na kraju ispisa su informacije o imenima varijabli u kojima je u aktivnom fajlu sačuvana klasterska pripadnost entiteta **/Variables (Cluster Membership) Saved into Working File/** ako je to odabrano u dijalogu **Save**.

U programu SPSS se grupisanje entiteta procedurom **K-means** izvodi odabirom menija **Statistics/Classify/K-Means Cluster** a potom se u dijalogu **K-Means Cluster Analysis** kao **Variables** unesu imena varijabli po kojima se grupišu entiteti a u polje **Label Cases by** može se (a ne mora) uneti ime tekstualne varijable koja definiše identifikaciju entiteta na ispisu. U polje **Number of Clusters** potrebno je uneti broj grupa koji želimo a u okviru **Method** definišemo da li želimo samo klasifikovanje entiteta u zadati broj grupa čiji će centri ostati tokom klasifikacije jednaki početnim centrima klastera (opcija **Classify**) ili klasifikovanje sa iteracijama (podešavanje centara klastera tokom pridruživanja novih entiteta- opcija **Iterate and Classify**). Ukoliko je odabran metod sa iteracijama u dijalogu **K-Means Cluster Analysis: Iterate** možemo (celim brojem od 1 do 999) definisati dozvoljen broj iteracija (**Maximum Iterations)** i (bilo kojim brojem od 0 do 1) kriterijum za konvergenciju rešenja (**Convergence Criterion**)[105] ili uključiti opciju **Use running means** koja dovodi do toga da se ažuriranje centara klastera vrši pri svakom pridruživanju novog entiteta u klaster. Ukoliko ova opcija nije uključena ažuriranje centara klastera vrši se tek pošto se svi entiteti razvrstaju u klastere.
U okviru dijaloga **K-Means Cluster Analysis: Options** možemo definisati statistike za ispis koje želimo a u okviru dijaloga **K-Means Cluster Analysis: Save** možemo zatražiti zapis u aktivnom fajlu s podacima varijabli koje sadrže klastersku pripadnost entiteta (**Cluster Membership**) i udaljenost entiteta od centra klastera kojem pripada (**Distance from Cluster Center**).
Početne centre klastera može definisati i sam korisnik tako što prethodno pripremi fajl sa centrima klastera u obliku fajla sa podacima. U fajlu sa definisanim centrima klastera prva varijabla ima ime **cluster_** a za svaki klaster u preostalim kolonama su vrednosti na varijablama prema kojima će se vršiti grupisanje. Varijable u ovim kolonama moraju imati ista imena kao i varijable prema kojima će biti vršeno grupisanje. U dijalogu **K-Means Cluster Analysis** treba kliknuti na dugme **Centers**, odabrati opciju **Read Initial from** (u okviru **Cluster centers**) i, posle klika na dugme **File**, upisati ime fajla (ili ga odabrati među ponuđenim fajlovima) koji sadrži početne centre klastera.
Ukoliko želimo da, posle izvršenog grupisanja, finalni centri klastera budu zapisani u nekom fajlu potrebno je da, polse klika na dugme **Centers** u dijalogu **K-Means Cluster Analysis**, u okviru **Cluster centers** uključimo opciju **Write final as**. Potom kliknemo na dugme **File** i upišemo ime fajla u kojem će biti sačuvani finalni centri dobijenih klastera.

[105]Kriterijum konvergencije je proporcija najmanje distance između početnih centara klastera. Iteracije prestaju kada potpuna iteracija ne menja centre klastera za distancu koja je veća od upisane proporcije kao kriterijuma konvergencije.

Ispis iz procedure **K-Means Cluster Analysis** sadrži:

-Pod naslovom **Q U I C K C L U S T E R** na početku ispisa su (ako je u okviru **Statistics** dijaloga **Options** odabrana opcija **Initial cluster centers)** početni centri klastera (tj.aritmetičke sredine grupa)/ **Initial Cluster Centers** a potom (ukoliko je rađena klasifikacija u iterativnom postupku) prikazane su promene u centrima klastera u pojedinim iteracijama (kolone **Iteration** i **Change in Cluster Centers**).
-Ako je u okviru **Statistics** dijaloga **Options** odabrana opcija **Cluster information for each case** pod naslovom **Case listing of Cluster membership** data je klasterska pripadnost entiteta (kolona **Cluster**) i udaljenost entiteta od centra klastera u koji je svrstan (kolona **Distance**).
-Pod naslovom **Final Cluster Centers** dati su konačni centri klastera (tj. aritmetičke sredine klastera na svim varijablama po kojima su entiteti grupisani) i matrica distanci među konačnim klasterima (**Distances between Final Cluster Centers**).
-Ukoliko je u okviru **Statistics** dijaloga **Options** odabrana opcija **ANOVA table** u ispisu se u odeljku **Analysis of Variance** prikazuju rezultati jednofaktorskih univarijatnih analiza varijansi u kojima je pripadnost klasteru faktor (nezavisna varijabla) a svaka varijabla koja je poslužila za klasterisanje tretirana je kao zavisna varijabla.
- Na kraju je dat pregled broja entiteta po klasterima (**Number of Cases in each Cluster**) i (ukoliko je to definisano u opciji **Save**) informacija o imenima varijabli u aktivnom fajlu sa podacima u kojima su sačuvane (**Variables Saved into Working File**) klasterska pripadnost entiteta **(Cluster Number)** i distance entiteta u odnosu na centar klastera u koji su svrstani **(Distance)**.

Primer:

Na uzorku izbeglica iz Hrvatske i Bosne i Hercegovine koje su utočište našle u našoj zemlji prikupljeni su, upitnikom za procenu ratnih stresora, podaci o stepenu izloženosti pojedinim kategorijama ratnih stresora. Bilo je neophodno, radi daljeg tretmana, grupisati ispitanike u sto homogenije grupe prema stepenu izloženosti sledećim kategorijama stresora:

1. borbeni;
2. fizička povreda;
3. stresori van linije fronta;
4. stresori u zarobljeništvu;
5. deprivacija;
6. gubitak vojne strukture.

Stepen izloženosti pojedinim kategorijama stresora iskazana je relativno kao broj doživljenih iskustava u odnosu na maksimalan broj iskustava koji su postojali u upitniku za datu kategoriju. Na taj način, skale izloženosti su se kretale od 0 do 1 pri čemu 1 znači maksimalni stepen izloženosti stresorima iz date kategorije.

ž* * * * * * * * * * * * * * P R O X I M I T I E S * * * * * * * * * * *

Data Information

195 unweighted cases accepted.
0 cases rejected because of missing value.

Squared Euclidean measure used.

- -

* * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * *

Agglomeration Schedule using Ward Method

Stage	Clusters Cluster 1	Combined Cluster 2	Coefficient	Stage Cluster 1st Appears Cluster 1	Cluster 2	Next Stage
1	109	148	.000000	0	0	10
2	143	146	.000000	0	0	3
3	116	143	.000000	0	2	5
4	126	141	.000000	0	0	7
5	116	136	.000000	3	0	57
6	65	130	.000000	0	0	15
7	44	126	.000000	0	4	8
8	44	121	.000000	7	0	41
9	48	112	.000000	0	0	18
10	5	109	.000000	0	1	16

----------Koraci od 11 do 189 su iseceni iz originalnog SPSS ispisa----------

190	18	172	13.511321	189	0	192
191	1	2	15.592070	188	177	194
192	16	18	18.743114	187	190	193
193	16	36	22.728853	192	186	194
194	1	16	38.877399	191	193	0

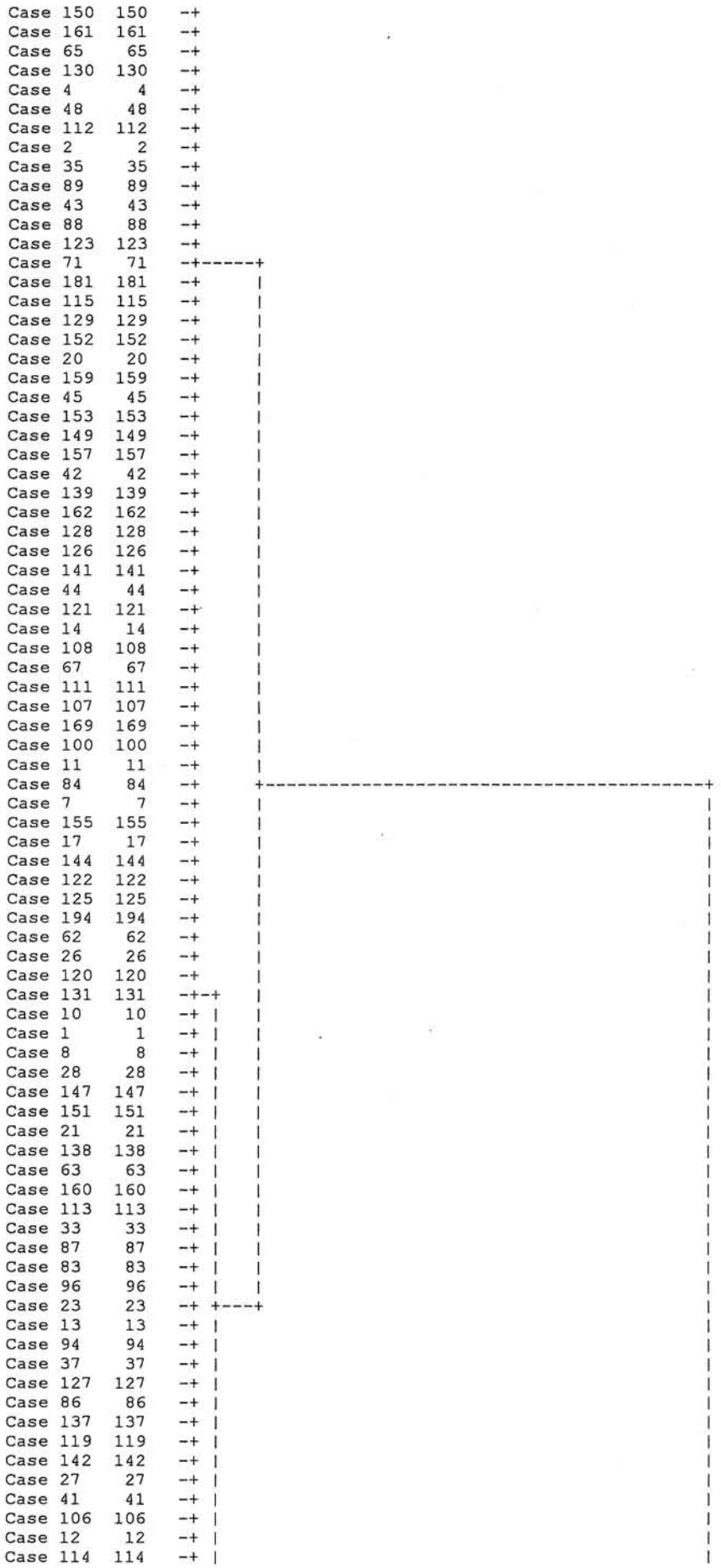
* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

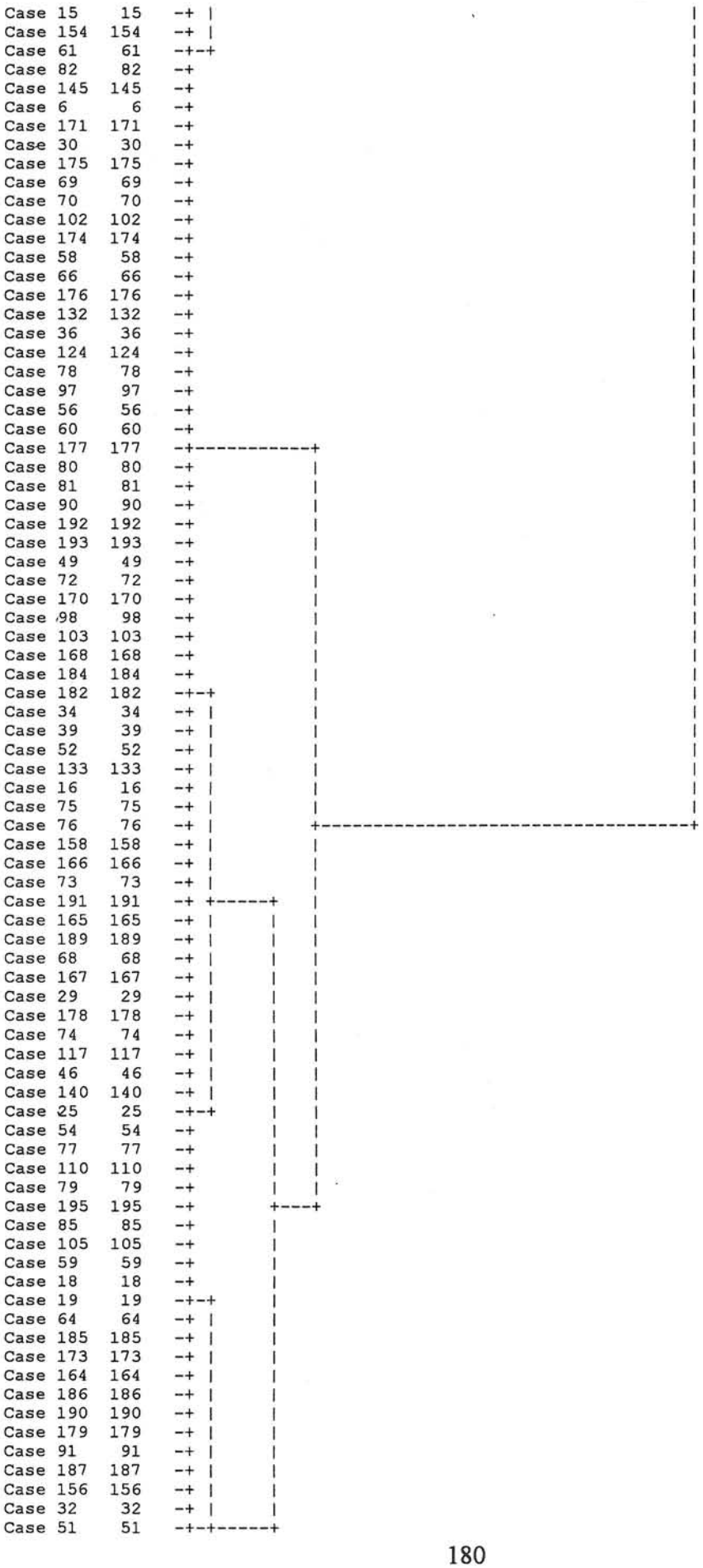
Dendrogram using Ward Method

Rescaled Distance Cluster Combine

```
   C A S E          0         5        10        15        20        25
 Label      Num    +---------+---------+---------+---------+---------+

 Case 109   109    -+
 Case 148   148    -+
 Case 5       5    -+
 Case 57     57    -+
 Case 93     93    -+
 Case 31     31    -+
 Case 55     55    -+
 Case 9       9    -+
 Case 24     24    -+
 Case 92     92    -+
 Case 53     53    -+
 Case 95     95    -+
 Case 38     38    -+
 Case 134   134    -+
 Case 135   135    -+
 Case 50     50    -+
 Case 163   163    -+
 Case 118   118    -+
 Case 143   143    -+
 Case 146   146    -+
 Case 116   116    -+
 Case 136   136    -+
 Case 22     22    -+
 Case 101   101    -+
 Case 3       3    -+
 Case 99     99    -+
 Case 40     40    -+
```


Case 150 150
Case 161 161
Case 65 65
Case 130 130
Case 4 4
Case 48 48
Case 112 112
Case 2 2
Case 35 35
Case 89 89
Case 43 43
Case 88 88
Case 123 123
Case 71 71
Case 181 181
Case 115 115
Case 129 129
Case 152 152
Case 20 20
Case 159 159
Case 45 45
Case 153 153
Case 149 149
Case 157 157
Case 42 42
Case 139 139
Case 162 162
Case 128 128
Case 126 126
Case 141 141
Case 44 44
Case 121 121
Case 14 14
Case 108 108
Case 67 67
Case 111 111
Case 107 107
Case 169 169
Case 100 100
Case 11 11
Case 84 84
Case 7 7
Case 155 155
Case 17 17
Case 144 144
Case 122 122
Case 125 125
Case 194 194
Case 62 62
Case 26 26
Case 120 120
Case 131 131
Case 10 10
Case 1 1
Case 8 8
Case 28 28
Case 147 147
Case 151 151
Case 21 21
Case 138 138
Case 63 63
Case 160 160
Case 113 113
Case 33 33
Case 87 87
Case 83 83
Case 96 96
Case 23 23
Case 13 13
Case 94 94
Case 37 37
Case 127 127
Case 86 86
Case 137 137
Case 119 119
Case 142 142
Case 27 27
Case 41 41
Case 106 106
Case 12 12
Case 114 114


Case 15 15
Case 154 154
Case 61 61
Case 82 82
Case 145 145
Case 6 6
Case 171 171
Case 30 30
Case 175 175
Case 69 69
Case 70 70
Case 102 102
Case 174 174
Case 58 58
Case 66 66
Case 176 176
Case 132 132
Case 36 36
Case 124 124
Case 78 78
Case 97 97
Case 56 56
Case 60 60
Case 177 177
Case 80 80
Case 81 81
Case 90 90
Case 192 192
Case 193 193
Case 49 49
Case 72 72
Case 170 170
Case 98 98
Case 103 103
Case 168 168
Case 184 184
Case 182 182
Case 34 34
Case 39 39
Case 52 52
Case 133 133
Case 16 16
Case 75 75
Case 76 76
Case 158 158
Case 166 166
Case 73 73
Case 191 191
Case 165 165
Case 189 189
Case 68 68
Case 167 167
Case 29 29
Case 178 178
Case 74 74
Case 117 117
Case 46 46
Case 140 140
Case 25 25
Case 54 54
Case 77 77
Case 110 110
Case 79 79
Case 195 195
Case 85 85
Case 105 105
Case 59 59
Case 18 18
Case 19 19
Case 64 64
Case 185 185
Case 173 173
Case 164 164
Case 186 186
Case 190 190
Case 179 179
Case 91 91
Case 187 187
Case 156 156
Case 32 32
Case 51 51

```
Case 104   104    -+ |
Case 183   183    -+ |
Case 180   180    -+ |
Case 188   188    -+ |
Case 47     47    -+ |
Case 172   172    ---+

* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Variables (Cluster Membership) Saved into Working File

    CLU5_1    to CLU2_1    for Ward Method
```

Iz dela ispisa naslovljenog **Agglomeration Schedule** vidimo da su u prvom koraku (**Stage 1**) spojeni ispitanici sa rednim brojevima 109 i 148 a iz kolone **Next Stage** da će im u 10 koraku biti pridružen ispitanik sa rednim brojem 5. U koloni **Stage cluster first appears** vidimo da se u trecem koraku (**Stage 3**) klasteru iz drugog koraka (u kojem su združeni ispitanici 116 i 143) pridružuje ispitanik broj 136. U koloni **Coefficients** možemo da pratimo porast distanci između klastera ili klastera i ispitanika koji se pridruzuje tom klasteru. U koraku 194, tj. pri pokusaju da se dva klastera spoje u jedan klaster) porast ovog koeficijenta sa 22.73 na 38.88 je nagliji nego dotada sto može da sugeriše da je rešenje sa bar dve grupe bolje od "rešenja" sa samo jednom grupom.

Dendrogram nam takođe sugeriše da je rešenje sa dve grupe sasvim zadovoljavajuće. Naime, vidimo da linija koja spaja dva poslednja klastera u jedan pokazuje nagli skok distance (sa ordinate čiji je naslov **Rescaled Distance Cluster Combine**) vidimo da je ovde skok u reskaliranoj distanci sa 8 (videti kod Case 75) na 25. Eventualno bi klaster koji je u donjem delu dendrograma mogao biti dalje razdvojen u dva.
U koju od grupa (za rešenja sa dva, tri, četiri i pet klastera je ispitanik svrstan prikazano je u nastavku ispisa a na kraju tog dela ispisa pod naslovom **CLU5_1 to CLU2_1 for Ward Method** program nas obaveštava o imenima varijabli u fajlu sa podacima u kojima je sačuvana informacija o klasterskoj pripadnosti ispitanika. Broj pre crtice u imenu varijable govori o tome o kojem rešenju (sa koliko klastera) je reč. Varijable u kojima je sačuvana klasterska pripadnost ispitanika mogu nam biti veoma korisne tokom odabira najboljeg rešenja i opisa karakteristika pojedinih klastera. Naime, ove varijable možemo upotrebiti da prebrojimo koliko ispitanika ima u svakom od klastera i da izračunamo prosečne vrednosti za ispitanike iz datog klastera varijablama na osnovu kojih smo izvršili grupisanje ili na nekim drugim karakteristikama ako ih imamo i da tako dobijemo profile koji će nam pomoći u tumačenju prirode pojedinih klastera.

Ukrštanjem pripadnosti klasterima (u rešenju sa dva klastera) sa borbenim statusom (učešće u borbenim operacijama pre izbegličtva), polnom pripadnošću i postojanjem dijagnoze postraumatskog stresnog poremećaja (PTSP) uočavamo da prvi klaster čine uglavnom žene i muškarci koji su u manjem stepenu imali učešća u borbenim operacijama. Isto tako, dijagnoza PTSP-a je relativno više zastupljena u drugom klasteru. Od 53 ispitanika koji imaju dijagnozu PTSP-a njih 36 je iz drugog klastera bez obzira na to što je taj klaster daleko manji po veličini.

STATUS status by CLU2_1 Ward Method

	CLU2_1		Page 1 of 1
Count			
			Row
	1	2	Total
STATUS			
1	95	11	106
izbeglica			54.4
2	30	59	89
izb.+ratnik			45.6
Column	125	70	195
Total	64.1	35.9	100.0

Number of Missing Observations: 0

POL by CLU2_1 Ward Method

	CLU2_1		Page 1 of 1
Count			
			Row
	1	2	Total
POL			
0	76	6	82
zene			42.1
1	49	64	113
muskarci			57.9
Column	125	70	195
Total	64.1	35.9	100.0

Number of Missing Observations: 0

DG14 ptsd by CLU2_1 Ward Method

	CLU2_1		Page 1 of 1
Count			
			Row
	1	2	Total
DG14			
0	108	34	142
ne			72.8
1	17	36	53
da			27.2
Column	125	70	195
Total	64.1	35.9	100.0

Number of Missing Observations: 0

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * *
* * * *

 Cluster Membership of Cases using Ward Method

                          Number of Clusters

  Label      Case       5     4     3     2

  Case 1        1       1     1     1     1
  Case 2        2       2     1     1     1
  Case 3        3       2     1     1     1

-Deo ispisa za ispitanike od 4 do 192 je isecen radi ustede prostora-

  Case 193    193       5     4     3     2
  Case 194    194       1     1     1     1
  Case 195    195       3     2     2     2

* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Variables (Cluster Membership) Saved into Working File

    CLU5_1    to CLU2_1    for Ward Method
```

Ispis iz klaster analize izvedene K-Means algoritmom (dva klastera):

```
* * * * * * * * * * * * Q U I C K     C L U S T E R * * * * * * * * * * * *

 Initial Cluster Centers.

  Cluster        E1              E2              E3              E4

        1          .0000           .0000           .0000           .0000
        2          .8667           .8889           .7000           .7778

  Cluster        E5              E6

        1          .0000           .0000
        2          .8667          1.0000

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Convergence achieved due to no or small distance change.
The maximum distance by which any center has changed is .0362
Current iteration is  4

Minimum distance between initial centers is 2.0946

Iteration        Change in Cluster Centers
                      1          2
     1               .3779      .8922
     2               .0487      .2553
     3               .0352      .1149
     4               .0202      .0629


 Final Cluster Centers.

  Cluster        E1              E2              E3              E4

        1          .0703           .0257           .1141           .0514
        2          .3979           .1250           .3104           .3480

  Cluster        E5              E6

        1          .2093           .0737
        2          .5618           .3490

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Analysis of Variance.

Variable	Cluster MS	DF	Error MS	DF	F	Prob
E1	3.8839	1	.019	193.0	195.3728	.000
E2	.3568	1	.014	193.0	24.5306	.000
E3	1.3951	1	.019	193.0	71.1049	.000
E4	3.1831	1	.021	193.0	148.6718	.000
E5	4.4964	1	.022	193.0	202.4660	.000
E6	2.7417	1	.020	193.0	133.2470	.000

Number of Cases in each Cluster.

Cluster	unweighted cases	weighted cases
1	147.0	147.0
2	48.0	48.0
Missing	0	
Valid cases	195.0	195.0

Variable Saved into Working File.

QCL_1 (Cluster Number)

Na početku ovog dela ispisa vidimo da su kao pocetni centri klastera (**Initial Cluster Centers**) uzeti ispitanici koji imaju najudaljenije rezultate na svim varijablama, tj. ispitanik BEZ ratnih traumatskih iskustava (na svim varijablama izlozenosti ratnim stresorima ima rezultat 0) uzet je kao početni centar prvog, a ispitanik sa skoro maksimalnom izloženošću ratnim stresorima kao početni centar drugog klastera. Proseci dva klastera na varijablama E1 do E6, na osnovu kojih smo vršili grupisanje dati su pod imenom Final cluster centers. Od 195 ispitanika njih 147 je svrstano u prvi a 48 u drugi klaster. Prvi klaster očigledno obuhvata ispitanike sa malim stepenom izloženosti svim kategorijama ratnih stresora dok drugi klaster obuhvata ispitanike koji su posebno bili izloženi borbenim, stresorima deprivacije i gubitku vojne strukture. Dva klastera se statistički značajno razlikuju u pogledu proseka na svim varijablama na osnovu kojih je vršeno grupisanje (Svi F-statistici u delu ispisa čiji naslov je **Analysis of variance** su statistički značajni). Pod imenom **QCL_1** u fajlu sa podacima upisana je varijabla u kojoj je sačuvana klasterska pripadnost ispitanika.

Ispis iz klaster analize izvedene K-Means algoritmom (rešenje sa tri klastera):

* * * * * * * * * * * * Q U I C K C L U S T E R * * * * * * * * * * * *

Initial Cluster Centers.

Cluster	E1	E2	E3	E4
1	.2000	.0000	.8000	.7407
2	.0000	.0000	.0000	.0000
3	.8667	.8889	.7000	.7778

Cluster	E5	E6
1	.6333	.0000
2	.0000	.0000
3	.8667	1.0000

Convergence achieved due to no or small distance change.
The maximum distance by which any center has changed is .0215
Current iteration is 7
Minimum distance between initial centers is 1.2766

Iteration	Change in Cluster Centers		
	1	2	3
1	.5613	.3082	.8343
2	.0788	.0409	.1447
3	.0622	.0263	.0538
4	.0445	.0173	.0498
5	.0282	.0092	.0623
6	.0321	.0058	.0726
7	.0303	.0144	.0000

Final Cluster Centers.

Cluster	E1	E2	E3	E4
1	.2537	.0584	.2621	.2781
2	.0339	.0187	.0975	.0258
3	.6137	.2418	.2706	.2810

Cluster	E5	E6
1	.4469	.2112
2	.1681	.0480
3	.6686	.5539

Analysis of Variance.

Variable	Cluster MS	DF	Error MS	DF	F	Prob
E1	2.9469	2	.009	192.0	309.7283	.000
E2	.3732	2	.012	192.0	29.6431	.000
E3	.6438	2	.020	192.0	31.7388	.000
E4	1.4835	2	.022	192.0	65.5016	.000
E5	2.8257	2	.016	192.0	173.2761	.000
E6	2.1094	2	.013	192.0	162.3947	.000

Number of Cases in each Cluster.

Cluster	unweighted cases	weighted cases
1	59.0	59.0
2	119.0	119.0
3	17.0	17.0
Missing	0	
Valid cases	195.0	195.0

Variable Saved into Working File.

QCL_2 (Cluster Number)

Izbeglice iz drugog klastera (njih 119) karakteriše nizak stepen izloženosti ratnim stresorima dok su izbeglice iz klastera 1 bile više izložene deprivaciji no ostalim vrstama ratnih stresora. Ispitanici iz treće grupe (njih 17) su bili u najvećem stepenu izloženi ratnim stresorima i to onima koji su karakteristični za učešće u borbenim operacijama (borbeni, deprivacija, gubitak vojne strukture). Da bismo preciznije otkrili sastav ovih klastera možemo da napravimo opet tabele kontingencije pripadnosti klasteru i nekih drugih karakteristika koje imamo o ispitanicima. U nastavku su date te tabele, dobijene procedurom **Statistics/Summarize/Crosstabs**, a u njima je prikazana povezanost klasterske pripadnosti sa jedne strane (u rešenjima sa dva i tri klastera) i polne pripadnosti, statusa (da li je reč o izbeglici koja je pre izbeglištva učestvovala u borbenim operacijama) i postojanja posttraumatskog stresnog poremećaja (varijabla DG14) sa druge strane.

Generalno, na osnovu sledećih nekoliko tabela uočljivo je da se žene i muškarci izbeglice svrstavaju pretežno u različite klastere. To je, pre svega,posledica toga što je izloženost

kategorijama ratnih stresora koji karakterišu učešće u borbenim dejstvima bila prisutnija kod muškaraca.

STATUS status by QCL_2

QCL_2 Page 1 of 1

		1	2	3	Row Total
STATUS					
1 izbeglica	Count	12	94		106
	Row %				54.4
2 izb.+ratnik	Count	47	25	17	89
	Row %				45.6
	Column Count	59	119	17	195
	Column %	30.3	61.0	8.7	100.0

Number of Missing Observations: 0

POL by QCL_2

QCL_2 Page 1 of 1

		1	2	3	Row Total
POL					
0 zene	Count	7	75		82
	Row %				42.1
1 muskarci	Count	52	44	17	113
	Row %				57.9
	Column Count	59	119	17	195
	Column %	30.3	61.0	8.7	100.0

Number of Missing Observations: 0

DG14 ptsd by QCL_2

QCL_2 Page 1 of 1

		1	2	3	Row Total
DG14					
0 ne	Count	32	105	5	142
	Row %				72.8
1 da	Count	27	14	12	53
	Row %				27.2
	Column Count	59	119	17	195
	Column %	30.3	61.0	8.7	100.0

Number of Missing Observations: 0

U sledećim dvema tabelama prikazan je stepen slaganja dveju primenjenih metoda u klasifikovanju ispitanika. Dok je slaganje u klasifikovanju ispitanika u dve grupe zavidno, pri klasifikaciji u tri klastera takvog slaganja uopšte nema. To je još jedan argument u prilog rešenju sa dva klastera.

QCL_1 by CLU2_1 Ward Method

QCL_1	CLU2_1 Count	1	2	Row Total
	1	125	22	147
				75.4
	2		48	48
				24.6
	Column Total	125	70	195
		64.1	35.9	100.0

Page 1 of 1

Number of Missing Observations: 0

QCL_2 by CLU3_1 Ward Method

QCL_2	CLU3_1 Count	1	2	3	Row Total
	1	7	39	13	59
					30.3
	2	118	1		119
					61.0
	3		17		17
					8.7
	Column Total	125	57	13	195
		64.1	29.2	6.7	100.0

Page 1 of 1

Koje od rešenja bismo prihvatili kao najbolje ponekad zavisi i od teorijskih i od praktičnih razmatranja. Ipak, na osnovu svega čini se da bi vodeći računa o izgledu dendrograma, ekonomičnosti (parsimoniji) i postojanju slaganja različitih metoda, rešenje sa dve grupe ovde bilo prihvatljivo. Reklo bi se da muškarci i žene u izbegliškoj populaciji čine različite subpopulacije u pogledu vrste ratnih stresora kojima su bili izloženi i stepena izloženosti ispitivanim kategorijama ratnih stresora. Ukoliko bismo uzeli u obzir rešenje sa tri klastera dobijeno k-means metodom (a koje se čini interpretabilnijim od rešenja sa tri grupe dobijenog Wordovom hijerarhijskom metodom) mogli bismo reći da u populaciji muškaraca izbeglica postoji subpopulacija koja je bila izložena visokom stepenu ratnih stresora koji su u vezi sa učešćem u borbenim operacijama i da je u toj subpopulaciji veći rizik nastanka PTSP-a nego u ostale dve subpopulacije.

XVI. KANONIČKA KORELACIONA ANALIZA

Kanonička korelaciona analiza je postupak za ispitivanje povezanosti dva skupa varijabli. Budući da svako od ova dva skupa izvornih, tj.merenih varijabli reprezentuje određenu oblast istraživanja, broj varijabli u svakom skupu može biti veliki. Ako bismo povezanost dva skupa varijabli ispitivali analizom pojedinačnih korelacija između varijabli jednog i drugog skupa bili bismo suočeni sa nepremostivim teškoćama:

1. Veoma je teško bilo šta zaključiti iz ogromnog broja pojedinačnih koeficijenata korelacije. Broj koeficijenata koji bismo morali uzeti u obzir, čak i ako bismo zanemarili korelacije među varijablama unutar svakog skupa, jednak je $m_1 * m_2$, tj. proizvodu broja varijabli u jednom i u drugom skupu.
2. Statistički bi bilo neispravno testirati statističku značajnost velikog broja koeficijenata korelacije jer ti statistički testovi ne bi bili nezavisni. Postojala bi šansa da, u velikom broju pojedinačnih korelacija, samo na osnovu slučaja određeni broj njih bude statistički značajnih.
3. Za analizu povezanosti dva skupa varijabli morali bismo uzeti u obzir i interkorelacije među varijablama unutar istog skupa.

Stoga se za analizu povezanosti dva skupa varijabli koriste postupci poput kanoničke korelacione analize koji daju globalniji pregled povezanosti varijabli iz dva skupa.

Struktura podataka

$E = \{ e_i ; i = 1, 2,...,n\} \subset P$ je slučajni uzorak izvučen iz homogene populacije;

$V_1 = \{ v_j ; j = 1, 2,..., m_1\} \subset U_1$ je prvi ("levi") skup kvantitativnih varijabli koji reprezentuje univerzum U_1;

$V_2 = \{ v_k ; k = 1,2,..., m_2\} \subset U_2$ je drugi ("desni") skup kvantitativnih varijabli koji reprezentuje univerzum U_2;

Pretpostavimo, dakle, da smo izmerili slučajni uzorak ispitanika, tj. skup E iz homogene populacije P u pogledu skupova V_1 i V_2, tj. u pogledu dva zasebna skupa kvantitativnih varijabli i da smo, potom, tako dobijene podatke standardizovali. Standardizovani rezultati na varijablama iz skupa V_1 su u matrici $\mathbf{Z}_1$, koja je reda n x m_1, a standardizovani rezultati na varijablama iz skupa V_2 su u matrici $\mathbf{Z}_2$, koja je reda n x m_2, ili jednostavnije rečeno:

$E \otimes V_1 = \mathbf{Z}_1 = (z_{ij})$; i = 1,...,n; j = 1,..., m_1.

$E \otimes V_2 = \mathbf{Z}_2 = (z_{ik})$; i = 1,...,n; k = 1,..., m_2.

Matrica $\mathbf{R}_{11} = \mathbf{Z}_1^t\mathbf{Z}_1 n^{-1}$ je tada matrica interkorelacija varijabli prvog skupa, $\mathbf{R}_{22} = \mathbf{Z}_2^t\mathbf{Z}_2 n^{-1}$ je matrica interkorelacija varijabli drugog skupa, a $\mathbf{R}_{12} = \mathbf{R}_{21}^t = \mathbf{Z}_1^t\mathbf{Z}_2 n^{-1}$ je matrica unakrsnih korelacija varijabli iz prvog sa varijablama iz drugog skupa.

Formalno definisanje problema:

$$\begin{array}{l|l} \mathbf{Z}_1\,\mathbf{x}_{1p} = \mathbf{k}_{1p} & 1.\ \mathbf{k}_{1p}{}^t\,\mathbf{k}_{2p}\,n^{-1} = \mathbf{x}_{1p}{}^t\,\mathbf{R}_{12}\,\mathbf{x}_{2p}\,n^{-1} = \text{maximum} \\ \mathbf{Z}_2\,\mathbf{x}_{2p} = \mathbf{k}_{2p} & 2.\ \mathbf{k}_{1p}{}^t\,\mathbf{k}_{1q}\,n^{-1} = \mathbf{k}_{2p}{}^t\,\mathbf{k}_{2q}\,n^{-1} = \delta_{pq} \\ & 3.\ \mathbf{k}_{1p}{}^t\,\mathbf{k}_{2q}\,n^{-1} = 0 \mid p \neq q \end{array}$$

Prvi indeks u prethodnoj notaciji je oznaka skupa kojem varijabla pripada, a drugi indeks je redni broj unutar istog skupa. To znači da je, na primer, k_{11} prva kanonička varijabla iz prvog skupa, k_{12} druga kanonička varijabla iz prvog skupa, a k_{21} prva kanonička varijabla iz drugog skupa. Kronekerova delta, u oznaci δ_{pq}, jednaka je 1 ako je p = q, a ako je $p \neq q$ onda je δ_{pq} jednaka 0.

Dakle, cilj kanoničke korelacione analize je napraviti linearne kombinacije izvornih varijabli unutar svakog skupa (kanoničke varijable $\mathbf{k}_{1p}$ i $\mathbf{k}_{2p}$) tako da:

1. kanoničke varijable iz različitih skupova ali sa istim redoslednim indeksom, tj indeksom p maksimalno koreliraju među sobom ;
2. kanoničke varijable budu standardizovane (ako je p=q Kronekerova delta δ_{pq} jednaka je 1, izrazi $\mathbf{k}_{1p}{}^t\,\mathbf{k}_{1q}\,n^{-1}$ i $\mathbf{k}_{2p}{}^t\,\mathbf{k}_{2q}\,n^{-1}$ su varijanse kanoničkih varijabli), a kanoničke varijable unutar istog skupa ne budu korelisane (akc je $p \neq q$ Kronekerova delta δ_{pq} jednaka je 0, a izrazi $\mathbf{k}_{1p}{}^t\,\mathbf{k}_{1q}\,n^{-1}$ i $\mathbf{k}_{2p}{}^t\,\mathbf{k}_{2q}\,n^{-1}$ su onda korelacije kanoničkih varijabli unutar istog skupa) i
3. kanoničke varijable prvog i drugog skupa koje imaju različit drugi indeks ne budu u korelaciji ($\mathbf{k}_{1p}{}^t\,\mathbf{k}_{2q}\,n^{-1} = 0 \mid p \neq q$).

Tako, druga kanonička varijabla iz prvog skupa ne sme korelisati ni sa jednom preostalom kanoničkom varijablom iz tog skupa, niti sa bilo kojom osim sa drugom, kanoničkom varijablom drugog skupa.

Problem se, zapravo, svodi na traženje jedinstvenog skupa vektora koeficijenata kojima treba pomnožiti rezultate na izvornim varijablam, tj. skupa vektora kanoničkih koeficijenata $\mathbf{x}_{1p}$ i $\mathbf{x}_{2p}$ koji će dati odgovarajuće linearne kombinacije, tj. proizvesti kanoničke varijable iz izvornih varijabli.

Parcijalnom derivacijom funkcije :

$$f(\mathbf{x}_{1p}, \mathbf{x}_{2p}, \lambda_{1p}, \lambda_{2p}) = \mathbf{k}_{1p}{}^t\,\mathbf{k}_{2p}\,n^{-1} - (1/2)\lambda_{1p}\,(1 - \mathbf{k}_{1p}{}^t\,\mathbf{k}_{1p}\,n^{-1}) - (1/2)\lambda_{2p}\,(1 - \mathbf{k}_{2p}{}^t\,\mathbf{k}_{2p}\,n^{-1}) =$$
$$= \mathbf{x}_{1p}{}^t\,\mathbf{R}_{12}\,\mathbf{x}_{2p}\,n^{-1} - (1/2)\lambda_{1p}\,(1 - \mathbf{x}_{1p}{}^t\,\mathbf{R}_{11}\,\mathbf{x}_{1p}\,n^{-1}) - (1/2)\lambda_{2p}\,(1 - \mathbf{x}_{2p}{}^t\,\mathbf{R}_{22}\,\mathbf{x}_{2p}\,n^{-1})$$

po $\mathbf{x}_{1p}$ i $\mathbf{x}_{2p}$ problem se svodi na rešavanje opšteg problema svojstvenih vrednosti:

$$(\mathbf{R}_{12}\,\mathbf{R}_{22}{}^{-1}\,\mathbf{R}_{21} - \lambda^2_{1p}\,\mathbf{R}_{11})\mathbf{x}_{1p} = 0 \Leftrightarrow (\mathbf{R}_{11}{}^{-1}\,\mathbf{R}_{12}\,\mathbf{R}_{22}{}^{-1}\,\mathbf{R}_{21} - \lambda^2_{1p}\,\mathbf{I})\,\mathbf{x}_{1p} = 0$$

ili

$$(\mathbf{R}_{21}\,\mathbf{R}_{11}{}^{-1}\,\mathbf{R}_{12} - \lambda^2_{2p}\,\mathbf{R}_{22})\,\mathbf{x}_{2p} = 0 \Leftrightarrow (\mathbf{R}_{22}{}^{-1}\,\mathbf{R}_{21}\,\mathbf{R}_{11}{}^{-1}\,\mathbf{R}_{12} - \lambda^2_{2p}\,\mathbf{I})\,\mathbf{x}_{2p} = 0$$

Vektor $\mathbf{x}_{2p}$ ukoliko znamo vektor $\mathbf{x}_{1p}$ možemo dobiti iz:

$$\mathbf{x}_{2p} = \mathbf{R}_{22}{}^{-1}\,\mathbf{R}_{21}\,\mathbf{x}_{1p}\,\lambda_{2p}{}^{-1}.$$

a vektor $\mathbf{x}_{1p}$ ukoliko znamo vektor $\mathbf{x}_{2p}$ možemo dobiti iz:

$$\mathbf{x}_{1p} = \mathbf{R}_{11}{}^{-1}\,\mathbf{R}_{12}\,\mathbf{x}_{2p}\,\lambda_{1p}{}^{-1}$$

Svojstvene vrednosti matrica $\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$ i $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ su jednake ($\lambda^2_{1p} = \lambda^2_{2p}$), a broj nenultih svojstvenih vrednosti (r) jednak je manjem od dva broja m_1 ili m_2, tj. broju izvornih varijabli u manjem skupu: r = minimum (m_1 , m_2). Prema tome, broj parova kanoničkih varijabli jednak je broju izvornih varijabli u manjem skupu.
Svojstvene vrednosti su, zapravo, kvadrati kanoničkih koeficijenata korelacije ($\lambda^2_{1p} = \lambda^2_{2p} = R_{cp}^2$), a svojstveni vektori $\mathbf{x}_{1p}$ i $\mathbf{x}_{2p}$ su traženi vektori kanoničkih koeficijenata.
Kanonička korelacija je u osnovi produkt-moment korelacija između kanoničkih varijabli, tj. između linearnih kombinacija izvornih varijabli jednog skupa i lineranih kombinacija izvornih varijabli drugog skupa, pri čemu linearne kombinacije iz različitih skupova imaju isti redosledni indeks. Kvadrat kanoničke korelacije odgovara zajedničkoj varijansi **kanoničkih** varijabli koje daju određenu korelaciju. Budući da je svaka sledeća svojstvena vrednost manja od prethodne, kao opšti pokazatelj kanoničke korelacije između dva skupa varijabli obično se prikazuje prva (i najveća) kanonička korelacija.

Testiranje hipoteza o populacionim kanoničkim koeficijentima korelacija:

H_0: $P_{cp} = 0$, (p-ta kanonička korelacija u populaciji jednaka je 0)

Pretpostavimo da izvorne varijable unutar svakog skupa imaju multivarijatnu normalnu funkciju gustine. (O multivarijatnoj funkciji gustine može se pročitati u knjizi Z. Kovačića Multivarijaciona analiza, str. 39 - 44). Tada se testiranje ove nulte hipoteze vrši sukcesivno za prvi (p=1) , drugi (p=2) i redom za p-ti koeficijent kanoničke korelacije računanjem statistika:
$\chi^2_p = -[n-1/2\,(m_1 + m_2 + 3)] \ln \Lambda_p$, pri čemu je Λ_p (Wilxova lamda) izraz

$$\Lambda_p = \prod_{p}^{r}(1 - R_{cp}^2) \text{ za } p = 1,2,\ldots,r.$$

($\prod_{p}^{r}$ označava proizvod izraza sa desne strane ovog znaka za vrednosti indeksa p od p do r, a R_{cp}^2 je oznaka p-tog kvadriranog koeficijenta kanoničke korelacije dobijenog na uzorku).
Wilxova Λ se može iskazati i na sledeći način:

$$\Lambda = |\, S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}\,| \;/\; |\, S_{22}\,|$$

pri čemu su S_{22}, S_{21}, S_{11} i S_{12} matrice kovarijansi dobijene na uzorku, izraz u brojiocu je determinanta proizvoda matrica, a u imeniocu se nalazi determinanta matrice S_{22}.
Statistik χ^2_p ima (ako je H_0 tačno) Hi-kvadrat distribuciju uzorkovanja sa stepenima slobode:

$$df = (m_1 - p + 1)\,(m_2 - p + 1).$$

Testiranje se zaustavlja kada se naiđe na prvu kanoničku korelaciju za koju se ne može odbaciti nulta hipoteza i sve ostale kanoničke korelacije koje slede za njom smatraju se statistički neznačajnim. Verovatnoća greške I tipa, tj. verovatnoća da je bar jedna od prvih k ($k \leq r$) kanoničkih korelacija pogrešno prihvaćena kao značajna, ako je H_0 tačno, je: $\alpha = 1-(1-\alpha_1)\,(1-\alpha_2)\,\ldots\,(1-\alpha_k)$. Na osnovu ovoga moguće je tako podešavati α_p da se dobije unapred fiksiran nivo rizika α. Može se, na primer, početi sa blažim nivoom rizika greške I tipa za prve kanoničke koeficijente a zatim taj nivo pooštravati. Pri tumačenju rezultata uzimaju se u obzir samo kanoničke varijable koje daju značajne kanoničke korelacije.

Matrice kanoničkih faktora (F_1 i F_2) i kanoničkih krosfaktora (C_1 i C_2):

U matricama kanoničkih faktora date su korelacije izvornih varijabli i kanoničkih varijabli istog skupa (opterećenja izvornih varijabli kanoničkim faktorima):

$$\mathbf{F}_1 = \mathbf{Z}_1^t \mathbf{K}_1 \; n^{-1} = \mathbf{R}_{11}\mathbf{X}_1$$

a

$$\mathbf{F}_2 = \mathbf{Z}_2^t \mathbf{K}_2 \; n^{-1} = \mathbf{R}_{22}\mathbf{X}_2$$

(Matrice $\mathbf{K}_1$ i $\mathbf{K}_2$ sadrže vektore kanoničkih varijabli a matrice $\mathbf{X}_1$ i $\mathbf{X}_2$ vektore kanoničkih koeficijenata).

U matricama kanoničkih krosfaktora nalaze se korelacije izvornih varijabli jednog skupa sa kanoničkim varijablama drugog skupa:

$$\mathbf{C}_1 = \mathbf{Z}_1^t \mathbf{K}_2 \; n^{-1} = \mathbf{R}_{12}\mathbf{X}_2$$

a

$$\mathbf{C}_2 = \mathbf{Z}_2^t \mathbf{K}_1 \; n^{-1} = \mathbf{R}_{21}\mathbf{X}_1$$

Matrice kanoničkih faktora i kanoničkih krosfaktora služe za tumačenje sadržaja kanoničkih varijabli i za analizu redundantnosti.

U slučaju da svaki skup izvornih varijabli sadrži samo po jednu varijablu, struktura $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ će biti jednaka r^2, a svojstvena vrednost skalara je jednaka tom skalaru. Kada kriterijski skup varijabli sadrži samo jednu a prediktorski skup više od jedne izvorne varijable onda se struktura $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ svodi na $\mathbf{q}^t\mathbf{R}^{-1}\mathbf{q} = \mathbf{b}^{*t}\mathbf{q} = R^2_{y.1,2,..m}$. Svojstvena vrednost od $R^2_{y.1,2,..m}$ jednaka je tom skalaru. Očigledno je da su Pirsonov produkt-moment koeficijent korelacije i multipla regresiona analiza samo specijalni slučajevi kanoničke korelacione analize.

Pri tumačenju ishoda kanoničke korelacione analize uzimaju se u obzir samo kanoničke varijable koje daju statistički značajne kanoničke korelacije.

Tumačenje kanoničkih varijabli na osnovu kanoničkih pondera, mada govori o relativnom doprinosu izvornih varijabli varijansi kanoničke varijable, radja mnoge probleme. Kanonički koeficijenti pridruženi pojedinim izvornim varijablama mogu biti niski ili čak negativni zbog korelacije ove varijable sa nekom drugom varijablom iz istog skupa pa je zato za tumačenje sadržaja kanoničkih varijabli bolje analizirati vektore kanoničkih faktora i krosfaktora.

Analiza prepokrivanja (redundantnosti)

Koeficijent redundantnosti

Kvadrat kanoničke korelacije govori o varijansi koju dele kanoničke a ne izvorne varijable iz dva skupa. Kao pokazatelj toga koliko ukupne varijanse izvornih varijabli jednog skupa je objašnjen (ili obuhvaćen) određenom kanoničkom varijablom iz tog skupa može se uzeti prosek kvadriranih korelacija izvornih varijabli i date kanoničke varijable.

Stuartov i Loveov koeficijent redundantnosti govori o procentu varijanse u kriterijskom skupu varijabli koji se može predvideti na osnovu prediktorskog skupa, tj. o delu varijanse jednog skupa koji je objašnjen kanoničkom varijablom drugog skupa:

Stjuart i Love tvrde da je koeficijent redundance jednak prosečnom koeficijentu multiple korelacije svake izvorne varijable jednog skupa sa optimalnim linearnim kombinacijama varijabli drugog skupa .
Koeficijent redundantnosti za skup y kada je dat skup x je:

$$R^2_{(j)y|x} = \lambda_j \sum_{j=1}^{m} \frac{(r^{(j)}_{y^*y_i})^2}{m_1} = \lambda_j R^2_{(j)y}$$

gde su λ_j kvadrati kanoničkih korelacija, $R^2_{(j)y}$ proporcije varijanse u skupu Y objašnjene pojedinom kanoničkom varijablom, m_1 broj varijabli u skupu a $r_{y^*y}{}^{(j)}$ vektor kanoničkih zasićenja za j-tu kanoničku varijablu.

Koeficijent redundantnosti za skup x kada je dat skup y je:

$$R^2_{(j)x|y} = \lambda_j \sum_{j=1}^{m} \frac{(r^{(j)}_{x^*x_i})^2}{m_2} = \lambda_j R^2_{(j)x}$$

Unakrsna zasićenja su proizvod kanoničke korelacije i kanoničkih zasićenja varijabli kanoničkom varijablom istog skupa.
Suma kvadriranih unakrsnih zasićenja datog skupa podeljena brojem varijabli u skupu daje koeficijent redundantnosti.
Stjuart-Loveova definicija redundantnosti izazvala je dosta rasprava i osporavanja tako da se sumnja u njenu matematičku korektnost.

Postoje i posebne statističke metode analize redundantnosti koje su definisali Van der Wollenberg (Redundancy analysis), DeSarbo (Canonical Redundancy Factoring Analysis) i Momirović sa saradnicima (Kanonička analiza kovarijanse u generalizovanom imaž prostoru).

Zapamtite:

- Pri primeni kanoničke korelacione analize korisno je napraviti **validaciju stabilnosti kanoničkih pondera i kanoničkih zasićenja**. To se može uraditi tako da se uzorak nasumice podeli u dva poduzorka, naprave kanoničke analize na svakom poduzorku i zatim koreliraju kanoničke varijable iz dveju analiza. Kanoničke koeficijente dobijene na jednom poduzorku možemo primeniti za računanje skorova na kanoničkim varijablama za drugi poduzorak i zatim izračunati korelacije tako dobijenih kanoničkih skorova (kanoničke korelacije). Ako su kanoničke varijable oba poduzorka u visokoj korelaciji i ako su kanoničke korelacije iz jednog poduzorka slične onima koje su dobijene na poduzorku čiji su koeficijenti korišćeni za računanje skorova na kanoničkim varijablama onda je pokazana stabilnost kanoničkih pondera i zasićenja.
- Pre konačnog tumačenja rezultata dobijenih kanoničkom korelacionom analizom potrebno je ustanoviti:
  Da li originalne, a ne samo kanoničke, varijable iz oba skupa imaju dovoljno veliki procenat zajedničke varijanse. To se može uraditi **analizom redundantnosti**;
  Ispitati postoje li vrlo visoke bivarijatne korelacije izmedu pojedinih varijabli iz jednog sa nekom izvornom varijablom drugog skupa. Postojanje malog broja takvih korelacija može dovesti do lažno visoke kanoničke korelacije.

- Pri primeni kanoničke korelacione analize treba voditi računa o tome da **uzorak** ispitanika bude bar **40 do 60 puta** veći od broja varijabli. Isto tako, prisustvo outliers-a, tj. ekstremnih rezultata može u velikoj meri uticati na veličinu kanoničke korelacije, pogotovu na malim uzorcima. Na kraju, **"slaba uslovljenost"** matrice interkorelacija izvornih varijabli istog skupa, do koje dolazi kada je jedna od varijabli linearna kombinacija ostalih varijabli iz datog skupa, onemogućava rešavanje problema ovom metodom. U tom slučaju nije moguće naći **regularni inverz** matrice interkorelacija varijabli datog skupa pa ni rešavanje **opšteg problema svojstvenih vrednosti** nije moguće.

****

Budući da u meniju SPSS-a nema mogućnosti da se direktno odabere kanonička korelaciona analiza ova analiza se u programu SPSS može uraditi uz pomoć odgovarajućeg makroa (npr.izvorni SPSS-ov makro **CANCORR.SPS** ili makroi domaćih autora /G.Knežević i K. Momirović/ **QCCR.SPS** i **KANONIKA.SPS**). Kanoničku korelacionu analizu je u SPSS programu moguće uraditi u okviru procedure za multivarijatnu analizu varijanse, ali je ispis iz ove procedure takav da je možda bolje koristiti za to odgovarajući makro. Ukoliko odgovarajući makro postoji u biblioteci makroa potrebno je otvoriti novi sintaksni prozor (**File/New/Spss Syntax**) i u njemu napisati komande:

INCLUDE '*ime makroa*.**SPS**'.
Ime makroa /**set1**=*imena prvog skupa varijabli*/**set2**=*imena drugog skupa varijabli* /**inc**=*nivo značajnosti*.
Poslednja mogućnost (**inc**=*nivo značajnosti*) je opciona samo u makrioma **QCCR.SPS** i **KANONIKA.SPS** i ukoliko se izostavi program će za testiranje značajnosti koristiti nivo 0.01.

Primer:
INCLUDE 'KANONIKA.SPS'.
KANONIKA SET1 = GRA1 to GRA26/**set2** =TWS to TABS/**inc**=0.05.

Prvi deo ispisa iz makroa **CANCORR.SPS**, **QCCR.SPS** i **KANONIKA.SPS** čine matrice interkorelacija varijabli unutar svakog skupa i matrica unakrsnih korelacija (korelacija varijabli iz jednog sa varijablama iz drugog skupa). Potom su dati koeficijenti kanoničke korelacije (sa testovima njihove značajnosti) i matrice kanoničkih koeficijenata (za makro **CANCORR.SPS** za sve parove kanoničkih varijabli dati su standardizovani i nestandardizovani (**Raw**) kanonički koeficijenti a za makroe **QCCR.SPS** i **KANONIKA.SPS** su dati standardizovani kanonički koeficijenti samo za one parove kanoničkih varijabli koje značajno koreliraju. U matricama strukture i krostrukture date su korelacije originalnih varijabli i kanoničkih varijabli unutar svakog skupa (kanonički faktori- Canonical loadings) kao i korelacije izvornih varijabli jednog skupa i kanoničkih varijabli drugog skupa (kanonički krosfaktori- Crossloadings). Na kraju su (samo ako se koriste makroi **QCCR.SPS** ili **CANCORR.SPS**) dati pokazatelji redundantnosti. U ispisu iz makroa **QCCR.SPS** analiza redundantnosti data je u odeljku Analiza prepokrivanja, a u ispisu iz makroa **CANCORR.SPS** u odeljku Redundancy Analysis.

Primer:
Na uzorku od 254 beogradskih srednjoškolaca muškog pola ispitivana je povezanost između samoocene sopstvene ličnosti i izraženosti psiholoških problema. (Prikazani primer nije zadovoljavajući u pogledu veličine uzorka i dat je samo kao ilustracija radi objašnjenja principa analize ispisa iz procedure).
Samoocena je u stvari stepen (ne)zadovoljstva pojedinim aspektima svoje ličnosti. Na upitniku samoocene ispitanici su na 7-stepenoj necifarskoj skali ocenjivali kojem od dva opozitna pola koji opisuju neku osobinu su oni bliži. Pored samoocene kakvi su po sopstvenoj proceni ispitanici su na isti način odgovorili kakvi bi želeli da budu.
Izraženost psiholoških problema, tj.problema u ponašanju merena je Ahenbahovim upitnikom za mlade. Postavlja se pitanje postoji li povezanost određenog složaja osobina sopstvene ličnosti kojima su mladi muškarci nezadovoljni i određenih složaja psiholoških problema koje imaju.
Prvi skup varijabli grade varijable koje pokazuju nezadovoljstvo pojedinim aspektima sopstvene ličnosti. Nezadovoljstvo pojedinim aspektima ličnosti iskazano je kao apsolutna razlika ocene kakvog sebe ispitanik smatra da jeste i kakav bi želeo da bude:

GRA1 APS.RAZL._IS VESEO_TUZAN
GRA2 APS.RAZL._IS NEPRAVEDAN_PRAVEDAN
GRA3 APS.RAZL._IS NESIGURAN_ SIGURAN
GRA4 APS.RAZL._IS ZANIMLJIV_NEZANIMLJIV
GRA5 APS.RAZL._IS LEP_RUZAN
GRA6 APS.RAZL._IS DOBAR·DRUG_LOS DRUG
GRA7 APS.RAZL._IS USAMLJEN_DRUŠTVEN
GRA8 APS.RAZL._IS ISKREN_NEISKREN
GRA9 APS.RAZL._IS PRICLJIV_ĆUTLJIV
GRA10 APS.RAZL._IS SAMOSTALAN_NESAMOSTALAN
GRA11 APS.RAZL._IS UPORAN_LAKO ODUSTAJEM
GRA12 APS.RAZL._IS NEODGOVORAN_ODGOVORAN
GRA13 APS.RAZL._IS TVRDOGLAV_POPUSTLJIV
GRA14 APS.RAZL._IS VERUJEM_NE VERUJEM
GRA15 APS.RAZL._IS LAKO OPRASTAM_TEŠKO OPRAŠTAM
GRA16 APS.RAZL._IS LAKO ODAJEM TAJNE-NE ODAJEM TAJNE
GRA17 APS.RAZL._IS TESKO SE LJUTIM-LAKO SE LJUTIM
GRA18 APS.RAZL._IS NAPET_OPUSTEN
GRA19 APS.RAZL._IS AGRESIVAN-NISAM AGRESIVAN
GRA20 APS.RAZL._IS OSETLJ. NA UVR-NEOSETLJIV NA UVREDE
GRA21 APS.RAZL._IS SLOBODAN-STIDLJIV
GRA22 APS.RAZL._IS NEZADOVOLJAN_ZADOVOLJAN
GRA23 APS.RAZL._IS SNALAZLJIV_NESNALAŽLJIV
GRA24 APS.RAZL._IS AMBICIOZAN_NEAMBICIOZAN
GRA25 APS.RAZL._IS NE POMAZEM_POMAZEM DRUGIMA
GRA26 APS.RAZL._IS SVADJAM SE CESTO-NE SVADJAM SE

Drugi skup varijabli čine varijable koje predstavljaju izraženost pojednih vrsta psiholoških problema, a izraženost problema je merena skalama problema u ponašanju sa ahenbahovog upitnika za mlade:

TWS WITHDRAWN SCALE-POVLAČENJE
TSCS SOMATIC COMPLAINTS SCALE-SOMATSKE TEŠKOĆE

TADS ANXIOUS/DEPRESSED SCALE-ANKSIOZNOST/DEPRESIVNOST
TSPS SOCIAL PROBLEMS SCALE-SOCIJALNI PROBLEMI
TTPS THOUGHT PROBLEMS SCALE-PROBLEMI U MIŠLJENJU
TAPS ATTENTION PROBLEMS SCALE-PROBLEMI PAŽNJE
TDBS DELINQUENT BEHAVIOR SCALE-DELINKVENTNO PONAŠANJE
TABS AGGRESSIVE BEHAVIOR SCALE-AGRESIVNO PONAŠANJE

Run MATRIX procedure:

Interkorelacije levog skupa varijabli
Columns 1 - 8

	GRA1	GRA2	GRA3	GRA4	GRA5	GRA6	GRA7
GRA1	1.000	.184	.261	.109	.151	.061	.295
GRA2	.184	1.000	.298	.183	.147	.285	.120
GRA3	.261	.298	1.000	.184	.141	.147	.213
GRA4	.109	.183	.184	1.000	.332	.154	.233
GRA5	.151	.147	.141	.332	1.000	.130	.140
GRA6	.061	.285	.147	.154	.130	1.000	.243
GRA7	.295	.120	.213	.233	.140	.243	1.000
GRA8	.163	.132	.161	.109	.073	.125	.048
GRA9	.064	.085	.043	.130	.158	.167	.192
GRA10	.151	.105	.188	.169	.153	.181	.163
GRA11	.178	.121	.161	.195	.055	.057	.094
GRA12	.130	.140	.083	.130	.130	.159	.098
GRA13	.139	.049	.034	.105	.113	.055	.104
GRA14	.125	.107	-.019	.135	.045	.213	.115
GRA15	.258	.019	-.003	.047	-.058	.089	.099
GRA16	.129	.172	.191	.140	.153	.119	.077
GRA17	.306	.048	.153	.134	.104	.070	.072
GRA18	.460	.153	.253	.108	.031	.166	.203
GRA19	.213	-.090	.044	.110	.021	.082	.203
GRA20	.223	.075	.198	.170	.001	-.042	.153
GRA21	.123	.086	.154	.113	.220	.103	.102
GRA22	.329	.237	.384	.265	.273	.213	.318
GRA23	.173	.198	.260	.216	.220	.234	.251
GRA24	.022	.130	.093	.230	.067	.238	.228
GRA25	.166	.103	.137	-.003	.060	.197	.184
GRA26	.122	.112	-.006	.130	.134	.069	.085

Columns 9 - 16

	GRA9	GRA10	GRA11	GRA12	GRA13	GRA14	GRA15
GRA1	.064	.151	.178	.130	.139	.125	.258
GRA2	.085	.105	.121	.140	.049	.107	.019
GRA3	.043	.188	.161	.083	.034	-.019	-.003
GRA4	.130	.169	.195	.130	.105	.135	.047
GRA5	.158	.153	.055	.130	.113	.045	-.058
GRA6	.167	.181	.057	.159	.055	.213	.089
GRA7	.192	.163	.094	.098	.104	.115	.099
GRA8	-.023	.152	.217	.173	.117	.096	.195
GRA9	1.000	-.007	-.045	.128	.129	.013	.098
GRA10	-.007	1.000	.223	.278	-.109	.105	.045
GRA11	-.045	.223	1.000	.245	-.017	.101	.153
GRA12	.128	.278	.245	1.000	.098	.043	.164
GRA13	.129	-.109	-.017	.098	1.000	.154	.234
GRA14	.013	.105	.101	.043	.154	1.000	.404
GRA15	.098	.045	.153	.164	.234	.404	1.000
GRA16	-.011	.148	.232	.269	.053	.096	.180
GRA17	.135	.028	.182	.083	.226	.146	.224
GRA18	.031	.122	.150	.096	.152	.165	.334
GRA19	.163	.148	.082	.078	.167	.255	.350
GRA20	.072	.154	.142	.084	.156	.113	.168
GRA21	.146	.084	.134	.016	.141	.113	-.029
GRA22	.144	.217	.169	.250	-.018	.080	-.004
GRA23	-.001	.339	.111	.116	-.011	.077	.014
GRA24	.115	.158	.318	.189	.093	.031	.062
GRA25	.077	.138	.020	.090	.056	.091	.116
GRA26	.227	.060	.042	.175	.176	.132	.185

Columns 17 - 24

	GRA17	GRA18	GRA19	GRA20	GRA21	GRA22	GRA23
GRA1	.306	.460	.213	.223	.123	.329	.173
GRA2	.048	.153	-.090	.075	.086	.237	.198
GRA3	.153	.253	.044	.198	.154	.384	.260
GRA4	.134	.108	.110	.170	.113	.265	.216

GRA5	.104	.031	.021	.001	.220	.273	.220
GRA6	.070	.166	.082	-.042	.103	.213	.234
GRA7	.072	.203	.203	.153	.102	.318	.251
GRA8	.243	.175	.209	.098	.164	.061	.144
GRA9	.135	.031	.163	.072	.146	.144	-.001
GRA10	.028	.122	.148	.154	.084	.217	.339
GRA11	.182	.150	.082	.142	.134	.169	.111
GRA12	.083	.096	.078	.084	.016	.250	.116
GRA13	.226	.152	.167	.156	.141	-.018	-.011
GRA14	.146	.165	.255	.113	.113	.080	.077
GRA15	.224	.334	.350	.168	-.029	-.004	.014
GRA16	.139	.160	.008	.116	.079	.205	.231
GRA17	1.000	.368	.195	.302	.211	.214	.087
GRA18	.368	1.000	.266	.319	.122	.321	.184
GRA19	.195	.266	1.000	.161	.100	.091	.110
GRA20	.302	.319	.161	1.000	.155	.181	.123
GRA21	.211	.122	.100	.155	1.000	.211	.200
GRA22	.214	.321	.091	.181	.211	1.000	.300
GRA23	.087	.184	.110	.123	.200	.300	1.000
GRA24	.098	.012	.004	.104	.078	.205	.172
GRA25	.002	.055	.127	-.035	.000	.111	.343
GRA26	.294	.123	.206	.161	.011	.019	-.026

Columns 25 - 26

	GRA25	GRA26
GRA1	.166	.122
GRA2	.103	.112
GRA3	.137	-.006
GRA4	-.003	.130
GRA5	.060	.134
GRA6	.197	.069
GRA7	.184	.085
GRA8	.133	.094
GRA9	.077	.227
GRA10	.138	.060
GRA11	.020	.042
GRA12	.090	.175
GRA13	.056	.176
GRA14	.091	.132
GRA15	.116	.185
GRA16	.190	.030
GRA17	.002	.294
GRA18	.055	.123
GRA19	.127	.206
GRA20	-.035	.161
GRA21	.000	.011
GRA22	.111	.019
GRA23	.343	-.026
GRA24	.041	.065
GRA25	1.000	-.040
GRA26	-.040	1.000

Interkorelacije desnog skupa varijabli

	TWS	TSCS	TADS	TSPS	TTPS	TAPS	TDBS
TWS	1.000	.392	.593	.577	.417	.451	.329
TSCS	.392	1.000	.486	.334	.420	.461	.383
TADS	.593	.486	1.000	.523	.547	.612	.368
TSPS	.577	.334	.523	1.000	.315	.554	.336
TTPS	.417	.420	.547	.315	1.000	.446	.364
TAPS	.451	.461	.612	.554	.446	1.000	.414
TDBS	.329	.383	.368	.336	.364	.414	1.000
TABS	.280	.379	.407	.340	.386	.547	.685

Kroskorelacije levog i desnog skupa varijabli

	TWS	TSCS	TADS	TSPS	TTPS	TAPS	TDBS
GRA1	.223	.238	.436	.154	.241	.272	.074
GRA2	-.007	.001	.063	.044	.003	.121	-.020
GRA3	.105	.120	.246	.159	.111	.207	.038
GRA4	.082	.019	.124	.153	-.010	.146	-.006
GRA5	.015	.030	.114	.132	.060	.183	.084
GRA6	.068	.131	.111	.049	.117	.073	-.043
GRA7	.387	.170	.383	.278	.242	.160	.074
GRA8	.052	.057	.103	.073	.100	.177	.122
GRA9	.238	.195	.161	.113	.103	.131	.212
GRA10	.041	.067	.134	.169	.051	.232	.066

GRA11	.044	.043	.129	.105	.079	.178	.059
GRA12	.097	.156	.066	.047	.111	.218	.182
GRA13	-.075	.076	.007	-.066	.004	-.026	.024
GRA14	.058	.068	.138	-.068	.053	.092	-.053
GRA15	.161	.171	.134	-.060	.139	.070	.059
GRA16	.070	.103	.060	.086	.128	.101	-.021
GRA17	.109	.225	.222	.027	.158	.140	.095
GRA18	.203	.179	.320	.054	.180	.203	-.019
GRA19	.173	.069	.209	.035	.046	.043	.058
GRA20	.110	.156	.175	.088	.046	.166	-.072
GRA21	.139	.053	.097	.142	-.035	.133	-.027
GRA22	.291	.159	.391	.202	.247	.289	.167
GRA23	.075	-.016	.113	.177	-.014	.157	-.002
GRA24	.098	.075	.101	.100	.041	.122	.035
GRA25	.075	.136	.099	.014	.181	.116	.066
GRA26	-.034	.119	.060	-.071	-.003	.035	.121

REZULTATI KANONICKE KORELACIJSKE ANALIZE

Koeficijenti kanonickih korelacija i njihova znacajnost:

	rho	rho2	lambda	hi2	df	sig
1	.632	.399	.210	367.942	208.000	.000
2	.513	.263	.349	248.075	175.000	.000
3	.441	.194	.473	176.076	144.000	.036
4	.404	.163	.588	125.175	115.000	.243
5	.362	.131	.702	83.280	88.000	.622
6	.321	.103	.808	50.279	63.000	.877
7	.243	.059	.900	24.727	40.000	.972
8	.207	.043	.957	10.354	19.000	.944

Kanonicki koeficijenti levog skupa varijabli

	1	2
GRA1	-.336	-.252
GRA2	.144	-.061
GRA3	-.051	-.181
GRA4	.061	-.205
GRA5	.071	-.059
GRA6	-.039	-.086
GRA7	-.462	-.214
GRA8	.011	.043
GRA9	-.150	.154
GRA10	.096	-.285
GRA11	.060	-.143
GRA12	.046	.119
GRA13	.216	.036
GRA14	-.114	.035
GRA15	-.103	.309
GRA16	-.008	.062
GRA17	-.112	.321
GRA18	-.182	.005
GRA19	-.093	.070
GRA20	-.046	-.210
GRA21	-.007	-.173
GRA22	-.323	.267
GRA23	.213	-.272
GRA24	-.040	.031
GRA25	-.058	.347
GRA26	.108	.215

Kanonicki koeficijenti desnog skupa varijabli

	1	2
TWS	-.493	.355
TSCS	-.101	.299
TADS	-.773	-.311
TSPS	.275	-.727
TTPS	-.140	.232
TAPS	.134	-.483
TDBS	.183	.020
TABS	.026	.773

Kanonicki faktori levog skupa varijabli

	1	2
GRA1	-.648	-.176
GRA2	-.032	-.195
GRA3	-.301	-.344
GRA4	-.119	-.322
GRA5	-.044	-.191

GRA6	-.214	-.099
GRA7	-.675	-.290
GRA8	-.086	.029
GRA9	-.291	.223
GRA10	-.075	-.424
GRA11	-.113	-.207
GRA12	-.082	.101
GRA13	.011	.174
GRA14	-.259	.114
GRA15	-.337	.396
GRA16	-.117	.026
GRA17	-.351	.272
GRA18	-.554	-.025
GRA19	-.366	.148
GRA20	-.283	-.217
GRA21	-.145	-.261
GRA22	-.580	-.104
GRA23	-.081	-.402
GRA24	-.139	-.097
GRA25	-.187	.201
GRA26	-.047	.300

Kanonicki faktori desnog skupa varijabli

	1	2
TWS	-.762	-.030
TSCS	-.495	.219
TADS	-.887	-.182
TSPS	-.346	-.510
TTPS	-.587	.197
TAPS	-.428	-.244
TDBS	-.188	.307
TABS	-.227	.451

Kanonicki kros-faktori levog skupa varijabli

	1	2
GRA1	-.409	-.091
GRA2	-.020	-.100
GRA3	-.190	-.177
GRA4	-.075	-.165
GRA5	-.028	-.098
GRA6	-.135	-.051
GRA7	-.426	-.149
GRA8	-.054	.015
GRA9	-.183	.115
GRA10	-.047	-.217
GRA11	-.072	-.106
GRA12	-.052	.052
GRA13	.007	.090
GRA14	-.164	.058
GRA15	-.213	.203
GRA16	-.074	.013
GRA17	-.222	.140
GRA18	-.350	-.013
GRA19	-.231	.076
GRA20	-.179	-.111
GRA21	-.091	-.134
GRA22	-.366	-.053
GRA23	-.051	-.207
GRA24	-.088	-.050
GRA25	-.118	.103
GRA26	-.030	.154

Kanonicki kros-faktori desnog skupa varijabli

	1	2
TWS	-.481	-.015
TSCS	-.313	.112
TADS	-.560	-.093
TSPS	-.219	-.262
TTPS	-.371	.101
TAPS	-.270	-.125
TDBS	-.119	.157
TABS	-.144	.231

------ END MATRIX -----

Dobili smo dve statistički značajne (na nivou 0.01 za koji smo se opredelili u ovom slučaju) kanoničke korelacije. To znači da između (ne)zadovoljstva sopstvenom ličnošću i postojanja određenih psiholoških problema postoji povezanost u populaciji mladih mušakarca. Kao što to vidimo iz kolone Sig u delu ispisa pod naslovom **REZULTATI KANONICKE KORELACIJSKE ANALIZE**, kada bismo se u zaključivanju odlučili za nivo 0.05 onda bi i treća kanonička korelacija bila značajna. Dakle, veliki broj pojedinačnih veza (u ovom sluaju 26*8) između zadovoljstva sopstvenom ličnossću i izraženosti određenih vrsta psiholoških problema možemo svesti na povezanost dveju struktura iz oba skupa. Kako su ove strukture, tj.kanoničke varijable nastale vidi se iz dela ispisa pod naslovom **Kanonicki koeficijenti desnog (levog) skupa varijabli**. Pregledom kolone 1 u delu ispisa pod naslovom **Kanonicki faktori levog (desnog) skupa varijabli** vidimo da psihički problemi koji spadaju u tzv. probleme internalizovanja (povlačenje, anksioznost&depresivnost, somatske teškoće, problemi mišljenja, problemi u socijalnoj komunikaciji) i koji čine prvu strukturu, tj. prvu kanoničku varijablu drugog skupa pokazuju povezanost sa nezadovoljstvom emocionalnim aspektima sopstvene ličnosti koji predstavljaju prvu strukturu iz prvog skupa i koji uključuju ne samo uobičajeno emotivno stanje (GRA1, GRA17, GRA18, GRA22) već i one aspekte koji su u vezi sa odnosom prema drugim ljudima (GRA3, GRA7, GRA9, GRA14, GRA15).

Druga kanonička korelacija predstavlja vezu između agresivnog i asocijalnog ponašanja (TABS i TDBS) koje ne uključuje lične probleme u komunikaciji sa drugima (TSPS ima negativan predznak u toj strukturi) sa jedne strane i (ne)zadovoljstva interpersonalnim aspektima sopstvene ličnosti (GRA10, GRA23, GRA26, GRA11, GRA7, GRA4, GRA3), sa druge strane.

Dakle, vrsta psiholoških problema koje imaju mladi muškarci, tj. da li su skloniji internalizovanju ili eksternalizovanju problema, u znatnoj meri se može nazreti iz toga kojim aspektima sopstvene ličnosti su nezadovoljni.

****

XVII. KANONIČKA DISKRIMINACIONA ANALIZA

Kanonička diskriminaciona analiza ima dva aspekta:

1. Ova metoda može poslužiti za utvrđivanje postojanja ili nepostojanja razlika između dveju ili više subpopulacija (definisanih kategorijama neke kategoričke varijable) u pogledu skupa kvantitativnih varijabli uzetih zajedno, tj. linearno kombinovanih u diskriminantnu funkciju. Ova namena diskriminacione analize obuhvata njen deskriptivni aspekt. Budući da niz aritmetičkih sredina na skupu kvantitativnih varijabli predstavlja vektor koji zovemo **centroid**, ovaj deskriptivni aspekt diskriminantne analize možemo definisati i kao metod kojim ispitujemo da li se subpopulacije, definisane kategoričkom varijablom, razlikuju u pogledu centroida na diskriminantnoj funkciji. Diskriminantna funkcija je linearna kombinacija skupa kvantitativnih varijabli oblika $D_i = b_0 + b_1x_1 + b_2x_2 + ...+ b_mx_m$, pri čemu se koeficijenti b_0 do b_m određuju tako da proseci grupa (centroidi) budu maksimalno međusobno udaljeni, a da, istovremeno, prosečna unutargrupna varijabilnost bude najmanja moguća.
2. Pored toga, diskriminaciona analiza se koristi i za klasifikovanje novih entiteta čija subpopulacijska pripadnost nije poznata, a za koje postoje rezultati na skupu kvantitativnih varijabli koje diskriminišu različite subpopulacije. Prema tome, pored deskriptivnog diskriminaciona analiza ima i klasifikacioni aspekt. Dakle, na osnovu rezultata na skupu kvantitativnih varijabli moguće je, korišćenjem diskriminantne funkcije, predvideti subpopulacijsku pripadnost entiteta za koji ne znamo kojoj subpopulaciji pripada.

U pogledu prvog aspekta diskriminaciona analiza je identična multivarijatnoj analizi varijanse, a u pogledu drugog svog aspekta ova metoda je jednaka regresionoj analizi u kojoj je kriterijumska varijabla kategorička.

Struktura podataka

$E = \{e_i;\ i = 1,\ldots, n\} \subset P = \bigcup_k^g P_k$ je slučajni uzorak izvučen iz populacije koja čini uniju g subpopulacija koje prirodno postoje ili su formirane eksperimentalno

$Q = \{ q_k ;\ k = 1,2,\ldots,g\}$ je kategorička varijabla koja se sastoji od g kategorija.

$V = \{ v_j ;\ j = 1,2,\ldots, m\} \subset$ U skup kvantitativnih varijabli

Merenjem jedinica uzorka E u pogledu skupa kvantitativnih varijabli V dobijamo matricu **X**:

$E \otimes V = \mathbf{X} = (x_{ij});\ i = 1,\ldots,n;\ j = 1,\ldots, m.$

Premda to nije nužno, pretpostavićemo da je izvršena standardizacija varijabli iz skupa V tako da su u matrici **Z** (skup kvantitativnih varijabli-V) podaci iz matrice **X** dati u standardizovanom obliku.

Subpopulacijsku pripadnost ispitanika kodiraćemo u kompletni disjunktni oblik:

$E \otimes Q = \mathbf{S} = (s_{ik});\ i = 1,\ldots,n;\ k = 1,\ldots, g.$

Matrica **S** je selektorska matrica dobijena tako da je:

$$s_{ik} = \begin{cases} 1, & e_i \in q_k \\ 0, & e_i \in q_k \end{cases} \quad i = 1, \cdots, n \quad k = 1, \cdots, r$$

Definišimo na osnovu podataka sledeće strukture:

$\mathbf{R} = \mathbf{Z}^t\mathbf{Z}n^{-1}$ je matrica interkorelacija varijabli iz skupa V

$\mathbf{G} = \mathbf{SM} = \mathbf{S}\,(\mathbf{S}^t\mathbf{S})^{-1}\,\mathbf{S}^t\,\mathbf{Z}$ je matrica reda nxm u kojoj su kao rezultati na nekoj varijabli v_j za svakog ispitanika upisane aritmetičke sredine na toj varijabli one grupe kojoj ispitanik pripada . Dakle, u toj matrici se nalaze rezultati ispitanika na varijablama iz V ali projektovani u g dimenzionalni hiperkub kojeg definišu binarne varijable iz **S**.

$\mathbf{M} = (\mathbf{S}^t\mathbf{S})^{-1}\,\mathbf{S}^t\,\mathbf{Z}$ je regresiona matrica reda gxm koja sadrži aritmetičke sredine varijabli po pojedinim grupama (i omogućuje da kada se predmnoži matricom **S** da se na osnovu grupne pripadnosti ispitanika predvidi njegov rezultat na nekoj varijabli sa najmanjom mogućom greškom).

Dakle,

$$\mathbf{SM} = \mathbf{Z} - \mathbf{E} \mid \mathrm{tr}\,(\mathbf{E}^t\mathbf{E}) = \min.$$

Odatle je

$$\mathrm{tr}\,(\mathbf{Z} - \mathbf{SM})^t\,(\mathbf{Z} - \mathbf{SM}) \text{ pa je } \mathbf{M} = (\mathbf{S}^t\mathbf{S})^{-1}\,\mathbf{S}^t\,\mathbf{Z}\,.$$

Matrica $(\mathbf{S}^t\mathbf{S})^{-1}$ dijagonalna matrica reda g x g čije dijagonalne vrednosti odgovaraju recipročnim vrednostima veličina grupa definisanih kategorijama varijable Q.

Definišimo i matrice suma kvadrata i unakrsnih prozvoda za varijable iz skupa V:

$$\mathbf{W} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ik} - \mathbf{m}_k)(\mathbf{x}_{ik} - \mathbf{m}_k)^t$$

$$\mathbf{B} = \sum_{k=1}^{g} n_k(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^t$$

$$\mathbf{T} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{x}_{ik} - \mathbf{m})(\mathbf{x}_{ik} - \mathbf{m})^t$$

Matrica **B** je matrica suma kvadrata i unakrsnih proizvoda između grupa a matrica **W** je matrica suma kvadrata i unakrsnih proizvoda unutar grupa. Matrica **T** je ukupna matrica suma kvadrata i unakrsnih proizvoda takva da je $\mathbf{T} = \mathbf{W} + \mathbf{B}$. Ove matrice su analogne sumama kvadrata odstupanja između i unutar grupa i ukupnoj sumi kvadriranih odtupanja kada je u skupu V samo jedna varijabla. Vektor $\mathbf{m}_{ik}$ je vektor aritmetičkih sredina na m varijabli za i-tog ispitanika, $\mathbf{m}_k$ je vektor aritmetičkih sredina k-te grupe a **m** je vektor aritmetičkih sredina za sve ispitanike zajedno.

Formalno definisanje problema:

Način prvi:

$\mathbf{Z}\,\mathbf{x}_p = \mathbf{k}_p$ | 1. $\mathbf{k}_p{}^t \mathbf{l}_p\, n^{-1} =$ maximum
$\mathbf{G}\mathbf{y}_p = \mathbf{l}_p$ | 2. $\mathbf{k}_p{}^t \mathbf{k}_q\, n^{-1} = \delta_{pq}$
3. $\mathbf{l}_p{}^t \mathbf{l}_q\, n^{-1} = \delta_{pq}$
4. $\mathbf{k}_p{}^t \mathbf{l}_q\, n^{-1} = 0 \mid p \neq q$

Pri tom je :

$\mathbf{k}_p{}^t \mathbf{l}_p\, n^{-1} = \mathbf{x}_p{}^t \mathbf{Z}^t \mathbf{G} \mathbf{y}_p\, n^{-1} = \mathbf{x}_p{}^t \mathbf{A} \mathbf{y}_p$

$\mathbf{k}_p{}^t \mathbf{k}_p\, n^{-1} = \mathbf{x}_p{}^t \mathbf{Z}^t \mathbf{Z} \mathbf{x}_p\, n^{-1} = \mathbf{x}_p{}^t \mathbf{R} \mathbf{x}_p$

$\mathbf{l}_p{}^t \mathbf{l}_p\, n^{-1} = \mathbf{y}_p{}^t \mathbf{G}^t \mathbf{G} \mathbf{y}_p\, n^{-1} = \mathbf{y}_p{}^t \mathbf{A} \mathbf{y}_p$

Matrica **A** je matrica korelacija varijabli iz *V* ali projektovanih u prostor kojeg definišu vektori matrice **S** (ili korelacija izvornih varijabli iz *V* i istih tih varijabli projektovanih u prostor kojeg definiše matrica **S**) dok je matrica **R** matrica izvornih varijabli iz skupa *V*. Formalno posmatrano, reč je o problemu koji je identičan problemu kanoničke korelacione analize. Zapravo, diksriminantna analiza se može posmatrati kao kanonička korelaciona analiza između skupa kvantitativnih varijabli , sa jedne, i jedne kategoričke, tj.skupa binarnih varijabli, sa druge strane. Dakle, potrebno je naći vektor $\mathbf{x}_p$ koji će omogućiti maksimalnu povezanost varijabli $\mathbf{k}_p$ i varijabli $\mathbf{l}_p$. Maksimalna povezanost varijabli $\mathbf{k}_p$ i varijabli $\mathbf{l}_p$ zapravo znači da treba naći vektor $\mathbf{x}_p$ takav da su na linearnoj kombinaciji izvornih varijabli iz *V* definisanoj kao $\mathbf{Z}\,\mathbf{x}_p$ (kanoničkoj diskriminacionoj funkciji ili varijabli) subpopulacije definisane varijablom Q što je moguće više udaljene, tj da centroidi grupa na diskriminacionoj funkciji budu maksimalno udaljeni. Pritom, različite linearne kombinacije varijabli iz *V* ne smeju međusobno biti u korelaciji.

Formalno rešenje postavljenog problema svodi se na rešavanje opšteg problema svojstvenih vrednosti:

$$(\mathbf{A} - \eta^2_p \mathbf{R})\,\mathbf{x}_p = 0$$

dakle na traženje svojstvenih vrednosti matrice **A** u prostoru kojeg definiše matrica **R** i njima odgovarajućih svojstvenih vektora, tj. nalaženje svojstvenih vrednosti i svojstvenih vektora nesimetrične matrice $\mathbf{R}^{-1}\,\mathbf{A}$.

Način drugi:

Treba maksimizovati količnik $\mathbf{v}_p{}^t \mathbf{B} \mathbf{v}_p / \mathbf{v}_p{}^t \mathbf{W} \mathbf{v}_p$, gde je $\mathbf{v}_p$ vektor pondera za diskriminantnu funkciju (linearnu kombinaciju varijabli iz skupa V). Ovaj problem se, tako, može posmatrati i kao problem maksimizovanja odnosa varijanse između grupa i varijanse unutar grupa na pojedinim diskriminantnim funkcijama. Rešenje ovako postavljenog problema uz uslov da je $\mathbf{v}_p{}^t \mathbf{W} \mathbf{v}_p = 1$ i da su sukcesivne diksriminacione funkcije nekorelisane svodi se na rešavanje problema:

$$(\mathbf{W}^{-1}\,\mathbf{B} - \lambda_p)\mathbf{v}_p = 0$$

Kvadrat kanoničke korelacije se potom može dobiti na osnovu svojstvenih vrednosti na sledeći način: $R^2_{cp} = \lambda_p/(1+\lambda_p)$. S druge strane, $\lambda_p = R^2_{cp} / (1- R^2_{cp})$.

Broj nenultih svojstvenih vrednosti, tj. broj linearnih kombinacija (kanoničkih diskriminacionih funkcija) koje se mogu dobiti (bez obzira na to na koji je od dva prikazana načina problem postavljen) odgovara manjem od dva broja: (g-1) ili m, pri čemu je g broj kategorija varijable Q, a m broj varijabli u skupu V. Ovo je posledica toga što je matrica $\mathbf{W}^{-1}\mathbf{B}$ iz koje izvlačimo svojstvene vrednosti proizvod dveju simetričnih matrica, pa je njen rang jednak rangu one matrice od ovih dveju koja ima manji rang. Treba uočiti takođe da, premda su diskriminantne funkcije međusobno nekorelisane vektori pondera za diskriminantne funkcije nisu međusobno ortogonalni. Dakle, između diskriminantnih funkcija postoji "statistička" ali ne i "geometrijska" ortogonalnost.
Naravno, najveći doprinos diskriminantnoj funkciji daće one izvorne varijable na kojima su aritmetičke sredine grupa najviše udaljene međusobno ili (što je isto) one varijable koje imaju najveću korelaciju sa kategoričkom varijablom Q.
Na taj način, odgovaramo na pitanje da li se subpopulacije definisanie kategorijama varijable Q međusobno razlikuju u pogledu skupa varijabli iz V i istovremeno dobijamo mogućnost (manje ili više uspešne) klasifikacije novih ispitanika na osnovu njihovih rezultata na varijablama iz V u jednu od kategorija varijable Q.

Matrica centroida grupa na p-toj diskriminantnoj funkciji($\mathbf{C}_p$):

$\mathbf{C}_p = \mathbf{MX}$ (gde je $\mathbf{X}$ matrica čije su kolone ponderi za diskriminantne funkcije) je matrica reda g x k čiji vektori su centroidi grupa na p diskriminantnih funkcija.

Centroidi grupa su aritmetičke sredine grupa na kanoničkoj diskriminacionoj funkciji. Centroid je zapravo tačka u m-dimenzionalnom prostoru čije su koordinate aritmetičke sredine grupe na m varijabli koje grade prostor.

Testiranje homogenosti matrica kovarijansi subpopulacija- Boxov test

Jedna od pretpostavki (osim multivarijatne normalne raspodele kvantitativnih varijabli iz skupa V u svakoj od subpopulacija) za korišćenje diskriminacione analize je i da su matrice kovarijansi varijabli iz skupa V za različite subpopulacije iste. Za testiranje zadovoljenosti ove pretpostavke Boxovim testom testiramo hipotezu:

H_0: $\Sigma_k = \Sigma \ \forall k$
Za testiranje H_0 o jednakosti matrica kovarijansi subpopulacija računa se, preko determinanti, tj. generalizovanih varijansi, M statistik:

$$M = \frac{\prod_{k=1}^{g} |\mathbf{S}_k|^{(n_k-1)/2}}{|\mathbf{S}|^{(n-g)/2}}$$

Matrica S (sa crticom iznad) u obrascu za M statistik je količnik W/(n-g), tj. ocena opšte kovarijacione matrice dok su matrice S_k ocene matrica kovarijansi u subpopulacijama. Na osnovu M statistika može se izračunati približni F-statistik i testirati H_0 na osnovu F-statistika:

Budući da je $-2b*\ln M \approx F$, gde je $b=[1-c_1-v_1/v_2]/v_1$

a

$$c_1 = \frac{(2m^2+3m-1)}{6(m+1)(g-1)}\left(\sum_{k=1}^{g}\frac{1}{n_k-1}-\frac{1}{n-g}\right)$$

$$c_2 = \frac{(m-1)(m+2)}{6(g-1)}\left(\sum_{k=1}^{g}\frac{1}{(n_k-1)^2}-\frac{1}{(n-g)^2}\right)$$

dok je

$v_1 = (1/2)m(m+1)(g-1)$

$v_2 = (v_1+2)\,|\,c_2-c_1^2|$

dobijeno F se distribuira po F-distribuciji sa $df_1=v_1$ i $df_2= v_2$.

U slučaju da je $c_2 - c_1^2$ manje od 0 količnik:

$$(-2b_1v_2\ln M)/(v_1 + 2\,b_1v_1\ln M),$$

pri čemu je

$$b_1 = [1- c_1 - (2/ v_2)]/ v_2),$$

se distribuira približno kao F-statistik sa $df_1=v_1$ i $df_2= v_2$.

Boxov test je veoma osetljiv i ne "reaguje" samo na nehomogenost matrica kovarijansi već i na odstupanje multivarijatne raspodele od normalne. Stoga je u slučaju da se njime odbaci hipoteza o jednakosti matrica kovarijansi u subpopulacijama često dovoljno da se klasifikacija u diskriminacionoj analizi vrši ne na osnovu zajedničke matrice kovarijansi unutar grupa (S sa crticom) već na osnovu odvojenih matrica kovarijansi za svaku grupu (S_k matrice).

<u>Testiranje hipoteza o populacionim kanoničkim koeficijentima korelacija</u>
<u>Utvrđivanje statističke značajnosti diskriminacionih funkcija:</u>

H_0: $P_{cp} = 0$ ili (što je isto) H_0: $\mu_{kp} = \mu$ za svako k.

Testiranje ove nulte hipoteze Bartletovim testom (pod uslovom da izvorne varijable unutar skupa *V* imaju multivarijatnu normalnu funkciju gustine) vrši se sukcesivno za prvi (p = 1), drugi (p = 2) i redom za p-ti koeficijent kanoničke korelacije računanjem statistika:

$$\chi^2_p = -[n-1/2\,(m+g+3)]\ln\Lambda_p,$$

gde je Λ_p (Wilxova lamda) izraz

$$\Lambda_p = \prod_{p}^{r} (1 / (1 + \lambda_p) \text{ za } p = 1,2..r.$$

Statistik χ^2_p se, ako je H_0 tačna, distribuira po Hi-kvadrat distribuciji a broj stepeni slobode jednak je (m - p + 1) (g - p).
Testiranje se zaustavlja kada se naiđe na prvu kanoničku korelaciju (tj. diskriminacionu funkciju) za koju se ne može odbaciti nulta hipoteza i sve ostale diskriminacione funkcije koje slede za njom smatraju se neznačajnim. Pri tumačenju rezultata uzimaju se u obzir samo one kanoničke diskriminacione funkcije koje daju značajne kanoničke korelacije.
Testiranje značajnosti diskriminacione funkcije identično je F-testu za testiranje statističke značajnosti razlika aritmetičkih sredina grupa na kanoničkoj diskriminativnoj funkciji.

Matrica strukture diskriminantnih funkcija (F) :

U matrici strukture diskriminantnih funkcije date su korelacije izvornih varijabli iz skupa *V* i kanoničkih diskriminacionih funkcija :

$$\mathbf{F} = \mathbf{Z}^t \mathbf{K}\, n^{-1} = \mathbf{RX}$$

Matrica **K** sadrži vektore rezultata na kanoničkim diskriminacionim funkcijama (diskriminacione skorove) a matrica **X** vektore koeficijenata za diskriminacione funkcije. Na osnovu vektora u matrici strukture koji odgovara datoj diskriminantnoj funkciji moguće je, na osnovu sadržaja izvornih varijabli koje sa tom funkcijom najviše koreliraju, odrediti sadržaj diskriminantne funkcije.

Klasifikacija entiteta

Najjednostavniji način da se novi entitet (čiju grupnu pripadnost ne znamo) klasifikuje u neku od grupa je da se izračuna kvadrirana euklidska distanca entiteta i centroida pojedinih grupa:

$$d_i^2 = (\mathbf{k}_{ip} - \mathbf{c}_{pk})^t (\mathbf{k}_{ip} - \mathbf{c}_{pk})$$

U obrascu za računanje distance, oznaka $\mathbf{k}_{ip}$ stoji za vektor rezultata ispitanika e_i na p-toj diskriminantnoj funkciji, a $\mathbf{c}_{pk}$ je vektor centroida grupe k na p-toj diskriminantnoj funkciji. Entitet (ispitanik) se potom svrstava u grupu za koju je njegovo d_i^2 najmanje.
Ovaj metod klasifikacije entiteta na uzima u obzir apriorne verovatnoće javljanja pojedinih subpopulacija. Pored ovoga, postoje i razrađeniji postupci klasifikovanja. Ukoliko se, pak, na osnovu datog uzorka želi ocena klasifikacione uspešnosti diskriminacione funkcije mogu se primeniti dva postupka:

1. Slučajnim odabirom podeliti uzorak na dva dela: na jednoj polovini uzorka se napravi diskriminaciona funkcija, a druga polovina uzorka se potom klasifikuje;
2. U svakom koraku izbaciti po jednog ispitanika iz računa kojim se pravi diskriminaciona funkcija, a zatim tog ispitanika klasifikovati. Ovaj se postupak često naziva "U-metod".

****

Kanonička diskriminaciona analiza se u SPSS programu radi odabirom menija **Statistics/Classify/Dscriminant**, tako što se u dijalogu **Discriminant Analysis** u polje **Independents** unesu imena kvantitativnih varijabli iz skupa V a u polje **Grouping Variable** ime kategoričke varijable Q i popune polja **Define Range** za tu varijablu. Korisno je pri tome odabrati u okviru **Descriptives** dijaloga **Discriminant Analysis: Statistics** opcije **Means**, **Box's M** i **Univariate ANOVAs**, kao i opciju **Fisher's** u okviru **Function Coefficients**. U dijalogu **Discriminant Analysis: Classification** svakako je potrebno odabrati **Summary Table** iz okvira **Display**, a u okviru **Prior Probabilities** default(podrazumevanu) opciju **All groups Equal** treba zameniti opcijom **Compute from group sizes** samo i samo tada kada je nejednaka verovatnoća pripadnosti subpopulacijama iz P dobro odražena u veličinama poduzoraka.

U novijim verzijama programa (SPSS 8.0), u okviru dijaloga **Discriminant Analysis: Classification**, može se zatražiti tzv. U-metod klasifikacije tako što se u okviru **Display** uključi opcija **Leave-one-out classification**. Program će u tom slučaju u svakom koraku izbaciti po jednog ispitanika iz postupka kojim se pravi diskriminaciona funkcija, a zatim tog ispitanika klasifikovati. Na taj tačin moguće je dobiti korektniju procenu klasifikacione uspešnosti diskriminantne funkcije. Međutim, ova opcija se može koristiti samo u slučaju kada se klasifikacija entiteta vrši na osnovu unutargrupne matrice kovarijansi, tj. sa uključenom opcijom **Within-groups** (u okviru **Use covariance matrix** dijaloga **Discriminant Analysis: Classification**). Ispis iz diskriminantne analize na početku sadrži veličine grupa (**Number of cases by group**) i (ukoliko je u okviru **Descriptives** dijaloga **Discriminant Analysis:Statistics** uključena opcija **Means**) aritmetičke sredine i standardne devijacije pojedinih grupa na pojedinačnim varijablama iz skupa V.

Pod naslovom **Wilks' Lambda (U-statistic) and univariate F-ratio** dati su (samo ako je u okviru **Discriminant Analysis:Statistics** uključena opcija **Univariate ANOVAs**) Wilxove lambde (koja su ovde jednake $1-\eta^2$ za svaku varijablu) i F-statistici identični onima iz procedure **One-Way ANOVA**.

U delu ispisa naslovljenom **Classification function coefficients** prikazani su Fisherovi koeficijenti koji služe za klasifikovanje ispitanika u jednu od kategorija tj. subpopulacija definisanih varijablom Q.

U odeljku ispisa **Canonical Discriminant Functions** date su svojstvene vrednosti (kolona **Eigenvalue**), procenat od ukupne međugrupne varijanse koji otpada na datu svojstvenu vrednost, tj. procenat ukupne diskriminacione moći skupa varijabli koji otpada na datu diskriminacionu funkciju (količnik date svojstvene vrednosti i zbira svih svojstvenih vrednosti pomnožen sa 100) u koloni **Pct of variance**, kanoničke korelacije (**Canonical Corr**) i rezultati testiranja značajnosti diskriminacionih funkcija (kanoničkih korelacija) pomoću Bartletovog Hi-kvdrat testa (kolone **Wilks' Lamda, Chi-square, df** i **Sig**). Testiranje značajnosti kanoničkih korelacija u suštini je isto što i testiranje

$$H_0: \mu_{kp} = \mu_p \ \forall k$$

(centroidi subpopulacija na p-toj diskriminantnoj funkciji su jednaki)

korišćenjem F-statistika u proceduri **One-Way ANOVA** u kojoj je p-ta diskriminaciona funkcija zavisna varijabla, a Q nezavisna varijabla, tj. faktor.

U nastavku ispisa dati su standardizovani (**Standardized canonical discriminant function coefficients**), nestandardizovani koeficijenti (samo ukoliko je u dijalogu **Discriminant Analysis:Statistics** izabrana opcija **Unstandardized** u okviru **Function Coefficients**) za pravljenje linearnih kombinacija, tj.diskriminacionih funkcija (**Unstandardized canonical discriminant function coefficients**), matrica korelacija izvornih varijabli i diskriminacionih funkcija (**Structure matrix**) kao i centroidi grupa na pojedinim discriminacionim funkcijama **/Canonical discriminant functions evaluated at group means (group centroids)/**. Standardizovani koeficijenti za diskriminantnu funkciju govore o parcijalnom doprinosu izvornih varijabli definisanju date diskriminantne funkcije. U delu ispisa koji nosi naslov **Test of Equality of Group Covariance Matrices Using Box's M** prikazani su rezultati testiranja homogenosti matrica kovarijansi subpopulacija. Ukoliko Boxov M statistik pokazuje nehomogenost matrica kovarijansi (vrednost u koloni **Significance** manja od 0.05) onda je bolje klasifikacioni aspekt diskriminacione analize napraviti sa uključenom opcijom **Separate- groups** (u okviru **Use covariance matrix** dijaloga **Discriminant Analysis: Classification**), a ne sa podrazumevanom opcijom **Within-groups**.
Ukoliko je u dijalogu **Discriminant Analysis: Classification** uključena opcija **Summary Table** iz okvira **Display** deo ispisa koji nosi naslov **Classification results** daje tabelu u kojoj su ukrštene stvarna (**Actual Group**) i predviđena (na osnovu diskriminantne funkcije-**Predicted Group Membership**) kategorijska pripadnost ispitanika. Procenat tačno klasifikovanih ispitanika (**Percent of "grouped" cases correctly classified**) je (pored visine kanoničke korelacije) mera uspešnosti diskriminantne funkcije u razlikovanju grupa na osnovu informacija sadržanih u skupu varijabli *V*.

****

Napomenimo ovde da je jednofaktorsku analizu varijanse moguće napraviti i diskriminacionom analizom, odabirom menija **Statistics/Classify/Dscriminant**, tj. napraviti je kao specijalni slučaj diskriminantne analize sa jednom kvantitativnom (polje **Independents!**) i jednom kategoričkom varijablom (polje **Grouping Variable**). Korisno je pri tome odabrati i u okviru **Descriptives** dijaloga **Discriminant Analysis:Statistics** opcije **Means** i **Univariate ANOVAs** kao i opciju **Summary Table** iz okvira **Display** dijaloga **Discriminant Analysis: Classification**. Ispis iz diskriminantne analize u kojoj je kategorička varijabla bila **Grouping Variable** a jedna zavisna kvantitativna varijabla predstavlja skup **Independent** (!) ima sve standardne komponente ispisa iz diskriminacione analize:
-Na početku ispisa date su veličine grupa (**Number of cases by group**) i (ukoliko je u okviru **Descriptives** dijaloga **Discriminant Analysis:Statistics** uključena opcija **Means**) aritmetičke sredine i standardne devijacije pojedinih grupa.
-Naravno, ovdě je koeficijent za kanoničku diskriminacionu funkciju (**Standardized canonical discriminant function coefficients**) jednak jedinici (jer samo jedna varijabla čini skup varijabli čijom linearnom kombinacijom treba dobiti diskriminacionu funkciju) a jedinici je jednaka i korelacija izvorne varijable sa samom sobom (tj. sa 'diskriminacionom funkcijom'-jedini element matrice strukture- **Structure matrix**) . Dakle, diskriminaciona funkcija je u ovom slučaju identična određenoj transformaciji izvorne kvantitativne varijable (zavisne varijable iz jednosmerne analize varijanse). Matrice kovarijansi su zapravo ukupne ili grupne ocene varijanse između i unutar grupa na toj varijabli.

Pod naslovom **Wilks' Lambda (U-statistic) and univariate F-ratio** dati su (samo ako je u okviru **Discriminant Analysis:Statistics** uključena opcija **Univariate ANOVAs**) Wilxova lambda (koja je ovde jednaka $1-\eta^2$) i F-statistik identičan onom iz procedure **One-Way ANOVA**.
U odeljku ispisa **Canonical Discriminant Functions** pod kanoničkom korelacijom (**Canonical Corr**) krije se zapravo η-koeficijent a testiranje njegove značajnosti pomoću Hi-kvadrat statistika (kolone **Wilks' Lamda, Chi-square, df** i **Sig**) isto je što i testiranje H_0 korišćenjem F-statistika u proceduri **One-Way ANOVA**.
S druge strane, ukoliko bismo jednofaktorsku analizu varijanse napravili sa diskriminantnim funkcijama, dobijenim u diskriminantnoj analizi, kao zavisnim varijablama, onda bi eta-koeficijenti bili jednaki kanoničkim korelacijama iz diskriminantne analize.

Primer1: Da li se u pogledu profila na poznatom instrumentu Minesota Multifazni inventar ličnosti razlikuju astmatičari od ostalih vrsta bolesnika? Na uzorku od 256 pacijenata, od kojih su polovina bili astmatičari (u varijabli grupa označeni jedinicom) a druga polovina bolesnici koji nemaju astmu (u varijabli grupa označeni dvojkom) izmerene su pomoću upitnika MMPI sledeće osobine ličnosti:

MMPIHs	HIPOHONDRIČNOST
MMPID	DEPRESIJA
MMPIHy	HISTERIJA
MMPIPd	PSIHOPATSKA DEVIJANTNOST
MMPIPa	PARANOIČNOST
MMPIPt	PSIHASTENIČNOST
MMPISc	SHIZOFRENIČNOST
MMPIMa	HIPOMANIČNOST
MMPIL	SKLONOST LAGANJU NA UPITNIKU
MMPIF	SKALA FALSIFIKOVANJU
MMPIK	KONTROLNA SKALA

Drugačije iskazano pitanje iz ovog primera moglo bi glasiti:Postoji li složaj ovih osobina ličnosti(diskriminantna funkcija) koji dobro diskriminiše astmatičare od ostalih bolesnika i, ako postoji, koliko uspešno na osnovu ddatog složaja možemo ubuduće vršiti predviđanja.

Ispis ove analize urađen je u verziji 6.1 programa SPSS. Budući da se ispis iz novijih verzija programa može razlikovati samo u tabeli koja prikazuje uspešnost klasifikacije ispitanika ako se koristi U-metoda, posle tumačenja ispisa za ovaj primer data je ova tabela iz verzije 8.0 pod naslovom "U-metoda klasifikovanja".

```
Following variables will be created upon successful completion of the
procedure:

  Name          Label
--------      ---------------------------------------
DIS1_1    --- Function 1 for analysis 1

*NOTE*  Option 'separate' means classification using group covariance matrices
        of the canonical discriminant functions, not those of the original
        variables.
        If there are fewer functions than variables, that makes a difference.
```

- - - - - - - - D I S C R I M I N A N T A N A L Y S I S - - - - - - - -

On groups defined by GRUPA

256 (Unweighted) cases were processed.
0 of these were excluded from the analysis.
256 (Unweighted) cases will be used in the analysis.

Number of cases by group

	Number of cases		
GRUPA	Unweighted	Weighted	Label
1	128	128.0	
2	128	128.0	
Total	256	256.0	

Group means

GRUPA	MMPIL	MMPIF	MMPIK	MMPIHS
1	3.95313	4.28125	11.89063	22.37500
2	3.71094	3.90625	12.03125	16.40625
Total	3.83203	4.09375	11.96094	19.39063

GRUPA	MMPID	MMPIHY	MMPIPD	MMPIPA
1	22.18750	21.03125	15.56250	12.25000
2	14.70313	16.79688	13.76563	8.41406
Total	18.44531	18.91406	14.66406	10.33203

GRUPA	MMPIPT	MMPISC	MMPIMA
1	25.45313	25.82813	13.18750
2	21.78125	24.69531	14.85938
Total	23.61719	25.26172	14.02344

Group standard deviations

GRUPA	MMPIL	MMPIF	MMPIK	MMPIHS
1	1.67830	2.17696	2.99668	4.63638
2	1.69368	1.85471	1.97997	2.28117
Total	1.68706	2.02702	2.53571	4.71582

GRUPA	MMPID	MMPIHY	MMPIPD	MMPIPA
1	5.25829	4.17329	4.05037	3.51823
2	3.26605	2.73893	3.06744	2.39266
Total	5.75692	4.11221	3.69690	3.56496

GRUPA	MMPIPT	MMPISC	MMPIMA
1	4.14284	4.42454	1.99112
2	2.84767	3.16616	2.60748
Total	3.99631	3.88131	2.46215

Wilks' Lambda (U-statistic) and univariate F-ratio
with 1 and 254 degrees of freedom

Variable	Wilks' Lambda	F	Significance
MMPIL	.99483	1.3206	.2516
MMPIF	.99141	2.2007	.1392
MMPIK	.99923	.1962	.6582
MMPIHS	.59794	170.7925	.0000
MMPID	.57580	187.1256	.0000
MMPIHY	.73389	92.1031	.0000
MMPIPD	.94071	16.0095	.0001
MMPIPA	.70941	104.0417	.0000
MMPIPT	.78812	68.2872	.0000
MMPISC	.97862	5.5490	.0193
MMPIMA	.88428	33.2401	.0000

- - - - - - - - D I S C R I M I N A N T A N A L Y S I S - - - - - -

On groups defined by GRUPA

Analysis number 1
Direct method: all variables passing the tolerance test are entered.

Minimum tolerance level..... .00100

Canonical Discriminant Functions
Maximum number of functions..... 1
Minimum cumulative percent of variance... 100.00
Maximum significance of Wilks' Lambda.... 1.0000
Prior probability for each group is .50000

Classification function coefficients
(Fisher's linear discriminant functions)

GRUPA =	1	2
MMPIL	1.3049984	.8367127
MMPIF	-.7902435	-.1793925
MMPIK	1.6418880	1.6849212
MMPIHS	.9687435	.5829751
MMPID	.3812024	.0835867
MMPIHY	.1866848	.0857148
MMPIPD	-.3442642	-.3283772
MMPIPA	.4052014	.0445016
MMPIPT	1.4639145	1.4335748
MMPISC	.7377774	.9388397
MMPIMA	2.7942773	3.1084807
(Constant)	-74.7584900	-66.3747971

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chi-sq.	df	Sig
					0	.375092	243.675	11	.0000
1*	1.6660	100.00	100.00	.7905					

* Marks the 1 canonical discriminant functions remaining in the analysis.

Standardized canonical discriminant function coefficients

	Func 1
MMPIL	.30705
MMPIF	-.48041
MMPIK	-.04250
MMPIHS	.54815
MMPID	.50660
MMPIHY	.13860
MMPIPD	-.02220
MMPIPA	.42202
MMPIPT	.04194
MMPISC	-.30082
MMPIMA	-.28347

Structure matrix:

Pooled within-groups correlations between discriminating variables and canonical discriminant functions

(Variables ordered by size of correlation within function)

	Func 1
MMPID	.66498
MMPIHS	.63530
MMPIPA	.49585
MMPIHY	.46653
MMPIPT	.40171
MMPIMA	-.28027
MMPIPD	.19451
MMPISC	.11451
MMPIF	.07212
MMPIL	.05586
MMPIK	-.02153

Unstandardized canonical discriminant function coefficients

	Func 1
MMPIL	.1821147
MMPIF	-.2375579
MMPIK	-.0167355
MMPIHS	.1500240
MMPID	.1157417
MMPIHY	.0392669
MMPIPD	-6.17839512E-03
MMPIPA	.1402749
MMPIPT	.0117990
MMPISC	-.0781924
MMPIMA	-.1221926
(Constant)	-3.2603894

Canonical discriminant functions evaluated at group means (group centroids)

Group	Func 1
1	1.28569
2	-1.28569

```
Test of Equality of Group Covariance Matrices Using Box's M

   The ranks and natural logarithms of determinants printed are those
   of the group covariance matrices.
     Group Label                    Rank   Log Determinant
         1                           11       23.849526
         2                           11       16.063686
      Pooled within-groups
      covariance matrix              11       21.685263

       Box's M  Approximate F  Degrees of freedom  Significance
     439.07879      6.35155     66,    205711.7        .0000

Classification results -

                        No. of    Predicted Group Membership
   Actual Group          Cases.         1          2
-------------------    ------    --------   --------

Group        1            128        110         18
                                   85.9%      14.1%

Group        2            128          4        124
                                    3.1%      96.9%

Percent of "grouped" cases correctly classified:  91.41%
Classification processing summary

       256 (Unweighted) cases were processed.
         0 cases were excluded for missing or out-of-range group codes.
         0 cases had at least one missing discriminating variable.
       256 (Unweighted) cases were used for printed output.
       256 cases were written into the working file.
```

Na početku ispisa, program nas obaveštava o tome da je u aktivnom fajlu sa podacima kreirana varijabla **DIS1_1** u kojoj su pohranjeni skorovi ispitanika na diskriminantnoj funkciji. Zatim su pod naslovima **Group means** i **Group standard deviations** dati proseci i standardne devijacije za pojedine grupe i uzorak u celini, na pojedinačnim zavisnim varijablama. Budući da je Box-ov M statistik od 439.08 (prikazan u ispisu pod naslovom **Test of Equality of Group Covariance Matrices Using Box's M**) statistički značajan i da su matrice kovarijansi između grupa heterogene klasifikaciju ispitanika na osnovu dobijene funkcije smo u ovom primeru uradili na osnovu zasebnih matrica kovarijansi unutar grupa (odabirom opcije **Separate- groups** u okviru **Use covariance matrix** dijaloga **Discriminant Analysis: Classification**),a ne sa podrazumevanom opcijom **Within-groups**.

Kao što vidimo u delu ispisa pod naslovom **Canonical Discriminant Functions** subpopulacija astmatičara se po svom profilu ličnosti razlikuje od neastmatičara. O tome govori statistički značajan Hi-kvadrat koji iznosi 243.675. To znači da između subpopulacijske pripadnosti (astmatičari-neastmatičari) sa jedne i skupa osobina ličnosti uzetih zajedno sa druge strane postoji povezanost u populaciji. Pod naslovima **Standardized canonical discriminant function coefficients** i **Unstandardized canonical discriminant function coefficients** možemo videti kako nastaje diskriminantna funkcija na osnovu standardizovanih, tj. nestandardizovanih rezultata na izvornim osobinama. Pored imena svake varijable dat je koeficijent kojim treba pomnožiti rezultat na datoj varijabli prilikom pravljenja linearne kombinacije, tj. diskriminantne funkcije.

Ipak, budući da su koeficijenti za diskriminantnu funkciju osetljivi na međusobne korelacije među sastavnim varijablama, za tumačenje sadržaja diskriminantne funkcije, tj. za definisanje složaja osobina koji najbolje diferencira astmatičare od ostalih bolesnika pogledaćemo koeficijente korelacije izvornih varijabli i diksriminantne funkcije koji su dati pod naslovom **Structure matrix**. Vidimo da je reč o složaju koji grade pre svega depresivnost, hipohondričnost, paranoidnost, histeričnost, i psihasteničnost. Ovu strukturu bismo možda najbolje opisali kao generalni neurotizam depresivno anksioznog tipa. Budući da je, kao što vidimo pod naslovom **Canonical discriminant functions evaluated at group means (group centroids)**,centroid, tj prosek astmatičara na dobijenoj diskriminatnoj funkciji pozitivan(tj. veći) možemo zaključiti i da su astmatičari generalno neurotičniji od ostalih bolesnika. Kvalitet diskriminantne funkcije, tj. diskriminativnost MMPI upitnika je, sudeći po visini kanoničke korelacije (0.79) i po rezultatima uspešnosti klasifikacije zadovoljavajuća. Iz tabele ispod naslova **Classification results** - vidimo da je 85.9% astmatičara klasifikovano tačno a da je 96.9% ostalih bolesnika klasifikovano u neastmatičare. Ipak, ne treba zaboraviti da smo klasifikovali iste ispitanike čiji su nam rezultati poslužili u pravljenju diskriminantne funkcije i da bi tek upotreba ove funkcije za klasifikovanje novih bolesnika dala pravu ocenu njene dikriminativnosti. U svakom slučaju, dobijeni rezulati sugerišu da postoji specifičan profil bolesnika sa astmom na MMPI instrumentu.

Ukoliko bismo sa dobijenom diskriminativnom funkcijom kao zavisnom varijablom i grupnom pripadnošću kao nezavisnom varijablom napravili jednofaktorsku univarijatnu analizu varijanse dobili bismo sledeći ispis:

- - Analysis of Variance - -

Dependent Variable DIS1_1 Function 1 for analysis 1
By levels of GRUPA

Value	Label	Mean	Std Dev	Sum of Sq	Cases
1		1.2856889	1.2229464	189.940927	128
2		-1.2856889	.7102128	64.0590734	128
Within Groups Total		-4.441E-16	1.0000000	254.000000	256

Source	Sum of Squares	d.f.	Mean Square	F	Sig.
Between Groups	423.1670	1	423.1670	423.1670	.0000
Within Groups	254.0000	254	1.0000		

Eta = .7905 Eta Squared = .6249

Uočimo da je eta koeficijent iz ove analize jednak kanoničkoj korelaciji iz diskriminantne funkcije.

U-metoda klasifikovanja

Tabela koja sledi deo je ispisa iz diskriminantne analize urađene u SPSS 8.0, a u kojoj je klasifikacija urađena na osnovu zajedničke unutargrupne matrice kovarijansi za obe grupe da bi se mogla primeniti "U - metoda" klasifikovanja. Tabela je prikazana samo da bi se ilustrovala razlika između starijih i novijih verzija SPSS-a u ovom delu ispisa, a ne kao deo

valjane analize u prethodno prikazanom primeru. Naime, budući da su matrice kovarijansi subpopulacija heterogene, u analizi koja je korišćena za Primer1 primenjena je valjana opcija za takav slučaj, tj. klasifikacija je urađena na osnovu odvojenih matrica kovarijansi za svaku grupu (opcija **Separate- groups** u okviru **Use covariance matrix** dijaloga **Discriminant Analysis: Classification**). Rezultati ove klasifikacije su već prikazani u prethodnom delu teksta i korišćeni su u tumačenju analize.

Classification Results

			Predicted Grou Membershi		Total
		GRUPA	Astm.	Neastm	
Origina	Count	Astm.	106	22	128
		Neastm	0	128	128
	%	Astm.	82,8	17,2	100,0
		Neastm.	,0	100,0	100,0
Cross validate	Count	Astm.	104	24	128
		Neastm.	2	126	128
	%	Astm.	81,3	18,8	100,0
		Neastm.	1,6	98,4	100,0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b 91,4% of original grouped cases correctly classified.
c 89,8% of cross-validated grouped cases correctly classified.

U tabeli treba uočiti dva dela (razdvojena vodoravnom zadebljanom linijom): u gornjem delu, pod nazivom **Original** je klasifikacija pomoću diskriminantne funkcije koja je konstruisana na osnovu rezulatat svih ispitanika a u donjem delu tabele pod nazivom **Cross-validated** je klasifikacija napravljena U-metodom. U ovom slučaju je od 128 astmatičara 106 njih je tako i klasifikovano običnom metodom a 104 U - metodom. U grupi neastmatičara tačno je klasifikovano svih 128 običnom metodom, a U - metodom njih 126.

Primer2: U ovom primeru biće prikazani i diskutovani samo delovi ispisa koji su specifični za situaciju kada ima više diskriminantnih funkcija, tj.više od dve grupe.

Na slučajnom reprezentativnom uzorku beogradskih srednjoškolki ispitivana je povezanost između potrebe za razgovorom sa stručanjakom za rešavanje psihičkih problema sa jedne i skupa varijabli koje opsuju stepen nezadovoljstva pojedinim aspektima sopstvene ličnosti. Način ispitivanja ovog nezadovoljstva opisan je u primeru iz odeljka XVI. Na osnovu načina odgovaranja na pitanje DA LI BI ŽELELA DA O SVOJIM PROBLEMIMA RAZGOVARAŠ SA STRUČNJAKOM ispitanice su pripale u jednu od tri kategorije: 0-NE ; 1-DA 2- MOŽDA.
Varijable koje definišu nezadovoljstvo pojedinim aspektima sopstvene ličnosti su varijable sa imenima od GRA1 do GRA26. Mere na ovim varijablama, kako smo već objasnili u odeljku XVI, su razlike između samoocena realnog ja (kakva sam ja sada) i idealnog ja (kakva bih želela da budem). Samoocene su davane odabirom odgovarajućeg kruga (od 3 sa leve i tri sa desne strane kvadrata koji se nalazio na sredini) koji je bliže odgovarajućem pridevu u

paru. Odabir kruga najbližeg društveno nepoželjnom ponašanju bodovan je ocenom 1 a odabir kruga najbližeg pridevu koji opisuje društveno poželjno ponašanje bodovan je ocenom 7. Veci rezultat na varijablama gra1 do gra26 znači manje zadovoljstvo (veću razliku realnog i idealnog ja).
Postoji li složaj pojedinih aspekata ličnosti koji dobro diskriminiše subpopulacije mladih devojaka definisane na osnovu odgovora na pitanje o potrebi za razgovorom sa stručnjakom.

```
Following variables will be created upon successful completion of the
procedure:


  Name          Label
--------     ----------------------------------------
DIS1_2    --- Function 1 for analysis 1
DIS2_2    --- Function 2 for analysis 1


 - - - - - - - -  D I S C R I M I N A N T  A N A L Y S I S  - - - - - -
On groups defined by RAZG      POTREBA ZA RAZGOVOROM SA STRUCNJAKOM
           205 (Unweighted) cases were processed.
             0 of these were excluded from the analysis.
           205 (Unweighted) cases will be used in the analysis.

Number of cases by group
                Number of cases
      RAZG  Unweighted     Weighted  Label
         0          112        112.0  NEMA POTREBU
         1           57         57.0  IMA POTR.
         2           36         36.0  MOZDA
     Total          205        205.0

 - - - - - - - -  D I S C R I M I N A N T  A N A L Y S I S  - - - - - -

On groups defined by RAZG      POTREBA ZA RAZGOVOROM SA STRUCNJAKOM
Analysis number       1
Direct method:  all variables passing the tolerance test are entered.
     Minimum tolerance level.....................  .00100
Canonical Discriminant Functions

     Maximum number of functions.....            2
     Minimum cumulative percent of variance...  100.00
     Maximum significance of Wilks' Lambda....  1.0000
Prior probabilities
   Group     Prior      Label

       0     .54634     NEMA POTREBU
       1     .27805     IMA POTR.
       2     .17561     MOZDA
   Total    1.00000

Classification function coefficients
(Fisher's linear discriminant functions)

RAZG     =          0              1              2
              NEMA POTREBU    IMA POTR.      MOZDA

GRA1              .6266518       .6224939       .6966435
GRA2              .5910929       .6896705       .7490775
GRA3              .0535577       .4585453       .3879950
GRA4              .5996213       .6134861       .5374034
GRA5              .2630857       .2019545       .2509061
GRA6             -.4565012      -.8873865     -1.0220751
```

GRA7	-.2230786	-.0892849	-.1289998
GRA8	.0731325	.1671920	.1245317
GRA9	.3858579	.6753808	.3423220
GRA10	.0302460	-.2261194	.1693685
GRA11	-.0960193	.0523014	-.0691201
GRA12	.0209326	.4154357	.3224807
GRA13	.5217866	.6526572	.4236102
GRA14	.1133608	.3085993	.2509504
GRA15	.5083442	.3978781	.3704967
GRA16	-.1356185	-.3001876	-.1719002
GRA17	.3079002	.3628076	.4266196
GRA18	.1305218	.3430499	.4767043
GRA19	-.0109233	.2100529	.0349709
GRA20	.1462887	-.0195954	-.1242902
GRA21	.0974052	-.2593994	-.0793181
GRA22	-.0527382	.2018748	.0730957
GRA23	.3282855	.4843992	.3096364
GRA24	.1649092	.0274038	-.0145707
GRA25	.2075344	-.0184537	-.0703716
GRA26	.3378757	.3239547	.2159675
(Constant)	-4.1531480	-7.4022728	-6.4197248

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chi-square	df	Sig
					: 0	.634435	86.226	52	.0020
1*	.4327	81.21	81.21	.5496	: 1	.908966	18.087	25	.8387
2*	.1002	18.79	100.00	.3017	:				

* Marks the 2 canonical discriminant functions remaining in the analysis.

Standardized canonical discriminant function coefficients

	Func 1	Func 2
GRA1	.01559	-.11821
GRA2	.10100	-.11727
GRA3	.48428	-.07603
GRA4	-.00525	.10503
GRA5	-.04573	-.04651
GRA6	-.31510	.26129
GRA7	.15001	.00613
GRA8	.06959	.02185
GRA9	.20419	.38851
GRA10	-.16927	-.54277
GRA11	.12985	.13730
GRA12	.46491	-.02769
GRA13	.08609	.35182
GRA14	.20482	.00783
GRA15	-.14013	.10494
GRA16	-.11271	-.10928
GRA17	.09982	-.16911
GRA18	.33334	-.39632
GRA19	.22055	.22010
GRA20	-.30473	.36303
GRA21	-.38988	-.15930
GRA22	.29764	.12221
GRA23	.11900	.21604
GRA24	-.15456	.12678
GRA25	-.25059	.18192
GRA26	-.05696	.23513

Structure matrix:

Pooled within-groups correlations between discriminating variables and canonical discriminant functions
(Variables ordered by size of correlation within function)

	Func 1	Func 2
GRA12	.52896*	.05717
GRA22	.50468*	.24766
GRA18	.45995*	-.11762
GRA3	.43246*	.11559
GRA1	.29232*	-.09320
GRA7	.28276*	.03571
GRA11	.22857*	.13201
GRA5	.21356*	.01378
GRA8	.19782*	.10958
GRA14	.19329*	.10312
GRA23	.16453*	.14968
GRA4	.15658*	.02186
GRA17	.12665*	-.08349
GRA16	-.04851*	.02294
GRA13	.09545	.41586*
GRA19	.23714	.39538*
GRA20	-.00403	.38864*
GRA9	.20792	.36942*
GRA26	.17852	.32721*
GRA25	-.08205	.31757*
GRA6	.05751	.26134*
GRA24	.12795	.22797*
GRA10	.13392	-.16597*
GRA15	-.07754	.16306*
GRA2	.07934	-.09860*
GRA21	-.02850	.08335*

* denotes largest absolute correlation between each variable and any discriminant function.

Unstandardized canonical discriminant function coefficients

	Func 1	Func 2
GRA1	.0110252	-.0835732
GRA2	.0850147	-.0987050
GRA3	.2892995	-.0454185
GRA4	-4.13732466E-03	.0827617
GRA5	-.0363679	-.0369835
GRA6	-.3477792	.2883943
GRA7	.0924470	3.77476629E-03
GRA8	.0621825	.0195245
GRA9	.1529484	.2910088
GRA10	-.1162622	-.3727922
GRA11	.0877328	.0927670
GRA12	.2771914	-.0165087
GRA13	.0541738	.2213868
GRA14	.1349615	5.15902748E-03
GRA15	-.0877976	.0657535
GRA16	-.0985705	-.0955635
GRA17	.0532088	-.0901421
GRA18	.1843500	-.2191849
GRA19	.1318181	.1315498
GRA20	-.1439616	.1715078
GRA21	-.2324032	-.0949566
GRA22	.1657887	.0680745
GRA23	.0833912	.1513914
GRA24	-.1107922	.0908799
GRA25	-.1788325	.1298261
GRA26	-.0309897	.1279344
(Constant)	-1.4901735	-.8143499

Canonical discriminant functions evaluated at group means (group centroids)

Group	Func 1	Func 2
0	-.58341	.05627
1	.85406	.29566
2	.46280	-.64320

Iz dela ispisa pod naslovom **Canonical Discriminant Functions** vidimo da je samo prva diskriminantna funkcija, tj. prva kanonička korelacija statistički značajna (Hi-kvadrat od 86.26 je za 52 stepena slobode statistički značajan jer je significance 0.002). Druga diskriminantna funkcija nije statistički značajna (Hi-kvadrat od 18.09 nije statistički značajan) i nju nećemo dalje tumačiti.

Na osnovu svega ovoga i izgleda kolone matrice strukture za prvu diskriminantnu funkciju (kolona **Funct 1** u ispisu naslovljenom **Structure matrix**) mogli bismo zaključiti da postoji povezanost između potrebe za razgovorom sa stručnjakom i izraženosti nezadovoljstva emocionalnim i interpersonalnim aspektima sopstvene ličnosti. Ta veza ipak nije jaka (prva kanonička korelacija je tek 0.55) . Na osnovu centroida na prvoj diksriminantnoj funkciji vidimo da je nezadovoljstvo onih koji IMAJU ili MOŽDA imaju potrebu za razgovorom sa stručnjakom veće no kod onih koji takvu potrebu nemaju.

****

XVIII. ANALIZA GLAVNIH KOMPONENTI

Cilj analize glavnih komponenti (engl. **Principal components analysis**) je transformisanje m izmerenih i međusobno povezanih varijabli u m ortogonalnih, tj. nekorelisanih glavnih komponenti. Svaka glavna komponenta je linearna kombinacija merenih, tj. izvornih varijabli. Dakle, u analizi glavnih komponenti broj glavnih komponenti teorijski, tj. matematički jednak je broju izvornih, merenih varijabli. To nije slučaj i sa metodama faktorske analize, pa je zbog ove suštinske konceptualne razlike bitno razlikovati analizu glavnih komponenti od faktorske analize. Dok faktorska analiza pretpostavlja određeni teorijski model o direktno nemerljivim latentnim "varijablama ili faktorima kao determinantama kovariranja (korelacija) merenih ili manifestnih varijabli, analiza glavnih komponenti najčešće rezultira svođenjem većeg broja merenih varijabli na manji broj glavnih komponenti ali bez teorijskih pretpostavki o kauzalnom uticaju komponenti na merene, opažene varijable.[106]

---

[106] Metode faktorske analize nisu obuhvaćene ovim Priručnikom. Budući da je veoma važno razlikovati analizu glavnih komponenti od metoda faktorske analize prenosimo deo teksta A. Zorića iz njegovog diplomskog rada u kojem su dobro objašnjeni osnovni principi faktorske analize:

"Faktorska analiza je multivarijantna statistička metoda koja služi otkrivanju unutrašnje strukture matrica podataka. Model faktorske analize se genuino smatra psihološkim...

Metodama faktorske analize nastoji se da se iz složene (multidimenzionalne) pojave, tj.skupa manifestnih obeležja, ekstrahuje manji broj važnih latentnih činilaca, tj. **faktora** koji su odgovorni za datu pojavu. Faktorskom analizom, dakle, pokušavamo da iza direktno izmerenih, manifestnih ili pojavnih varijabli koje smo izmerili na nekom reprezentativnom uzorku entiteta, i uz pomoć kojih nastojimo da opišemo izučavanu pojavu, pronađemo manji broj latentnih, skrivenih varijabli koje nazivamo **faktorima**, a koje objašnjavaju posmatranu pojavu, tj. čijom se kombinacijom mogu dobiti pojavne varijable. Kako o pojavi koju proučavamo saznajemo samo preko varijabli koje predstavljaju mere pojedinih aspekata te pojave (iz ovoga sledi i posledica da naše varijable moraju biti reprezentativan skup svih varijabli koje mere datu pojavu, tj. da naš skup varijabli mora dobro, adekvatno opisivati izučavanu pojavu), to je očigledno da termin "objasniti pojavu" znači naći odgovarajuću matematičku vezu između izmerenih varijabli, tj. objasniti dobijene veze između njih, ili pronaći uzroke pojave, tj. objasniti dobijenu varijansu izmerenih varijabli. Na osnovu ove razlike u pristupu objašnjenju izučavane pojave, faktorske metode možemo podeliti u dve velike kategorije:

- **faktorske metode u užem smislu** - metode kojima pokušavamo da objasnimo veze, tj. korelacije između manifestnih, izmerenih varijabli;
- **komponentne metode** - postupci kojima nastojimo da objasnimo variranja u merama izučavane pojave iznalaženjem njenih uzroka.

Faktorske metode u užem smislu nastoje da objasne veze među posmatranim varijablama. One polaze od pretpostavke da su manifesne, direktno izmerene varijable međusobno povezane, dakle u korelaciji, ne zato što su one u nekom uzročno - posledičnom odnosu, već zato što ih sve određuju neki njima zajednički faktori. Na ovaj način, cilj ove grupe metoda postaje rekonstrukcija korelacione matrice, dakle veza među varijablama preko manjeg broja zajedničkih faktora. U literaturi se zbog ovoga faktorski modeli u užem smislu nazivaju i analizom zajedničkih faktora.

U ovom modelu varijansa svake direktno merene, tj. manifestne varijable se razbija na dva dela:

- deo koji je zajednički sa drugim varijablama na osnovu kojega one koreliraju. Ovaj deo varijanse objašnjavaju zajednički faktori. Varijable koreliraju međusobno zato što koreliraju sa **zajedničkim** faktorima, te se ovaj deo varijanse varijable zove **komunalitet** i obeležava se sa h^2;
- deo varijanse varijable koji ne objašnjavaju zajednički faktori, te koji je zbog toga samo njoj svojstven, i koji objašnjava njoj odgovarajuća, takođe latentna varijabla koja se naziva **unikni** faktor. Jasno je da po ovoj podeli unikni faktor može biti povezan samo sa jednom odgovarajućom varijablom, jer bi u suprotnom i on predstavljao uzrok zajedničkog variranja varijabli, a samim tim bi po definiciji bio zajednički faktor. Ovaj deo varijanse se naziva unikvititet i obeležava se sa u^2. Unikna varijansa se dalje deli na dva dela: deo prave varijanse koji je specifičan za datu varijablu i koja se zato zove **specificitet**, a obeležava sa s^2, i na deo koji je posledica **greške** i koji se obeležava sa e^2.

Komponentne metode faktorske analize nastoje da pronađu latentnu strukturu koja stoji iza direktno izmerenih varijabli i koja je odgovorna za dobijena variranja na tim varijablama. Komponentne metode dakle

U analizi glavnih komponeti se ne pravi distinkcija između zajedničke i unikne varijanse merenih (ili faktorskim jezikom rečeno manifestnih) varijabli kao što je to slučaj u faktorskom modelu.[107] Sve glavne komponente obuhvataju celokupnu varijansu izvornih, merenih varijabli, ali tako da na svaku sledeću glavnu komponentu otpada sve manji deo ove ukupne varijanse. Ipak, praktični cilj ove analize je da se, od svih m glavnih komponeti koje se dobijaju na osnovu m izvornih, tj. izmerenih varijabli zadrži samo mali broj komponenti u kojima će pri tom, biti sadržana što veći deo ukupne varijanse (ili količine informacija) svih izvornih, tj. izmerenih varijabli. Budući da ova analiza ima smisla samo ako su izvorne varijable međusobno korelisane[108], tj. bar u izvesnoj meri redundantne analiza glavnih komponenti najčešće rezultira zadržavanjem manjeg broja komponenti koje se koriste u daljim analizama, npr. kao prediktorske varijable u regresionoj analizi.
Analiza glavnih komponenti nije metrički invarijatna: glavne komponente dobijene na osnovu matrice kovarijansi, tj. iz nestandardizovanih izvornih varijabli ne moraju odgovarati komponentama koje se dobijaju iz matrice interkorelacija, tj. na osnovu standardizovanih izvornih varijabli.[109]

Struktura podataka

$E = \{ e_i ; i = 1,...,n\} \subset P$ je slučajni uzorak izvučen iz homogene populacije;

$V = \{ v_j ; j = 1,2,..., m\} \subset U$ je skup kvantitativnih varijabli koji reprezentuje univerzum U svih mogućih varijabli koje se odnose na pojavu koja se proučava;

Pretpostavićemo da su podaci standardizovani, što je, s obzirom na osetljivost, tj. neinvarijantnost analize glavnih komponenti na metriku (tj. merne jedinice ili skalu varijable), preporučljiv postupak u psihološkim istraživanjima. Naime, s obzirom da su metrika (nulta, početna tačka i skala) za skoro sve psihološke varijable arbitrarni najčešće nema mnogo smisla raditi analizu glavnih komponenti na nestandardizovanim varijablama.

---

nastoje da pronađu uzroke pojave, koju registrujemo preko variranja rezultata u merenjima koja vršimo nad njom. Pod uzrokom se onda podrazumeva latentna varijabla, faktor ili u ovom slučaju komponenta koja je odgovorna za nastalu varijansu u rezultatima merenja. Za ovu varijansu može biti odgovorna i greška koja nastaje prilikom merenja, te drugi problem u komponentnim metodama jeste i određivanje da li izolovana latentna varijabla predstavlja realno egzistirajuću strukturu ili predstavlja samo grešku..." (A. Zorić, Faktorska analiza binarnih varijabli u generalizovanom imaž prostoru, Diplomski rad, Beograd:Filozofski fakultet, 1996.).

[107] Ključna razlika ovog postupka u odnosu na **komponentni model faktorske analize** je u tome što se faktorska analiza izvodi na matrici interkorelacija pa je, prema tome, metrički invarijatna i u tome što se u faktorskoj analizi zadržavaju samo neke a ne sve glavne komponente kao faktori. Komponentni model faktorske analize je zapravo korišćenje analize glavnih komponenti za aproksimaciju modela faktorske analize u užem smislu, tj. za aproksimaciju zajedničkih faktora.

[108] Provera statističke hipoteze: **P = I** (matrica interkorelacija varijabli u populaciji iz koje je uzorak jednaka je matrici identiteta, tj. varijable u populaciji nisu linearno povezane) zove se **Bartletov test sferičnosti**. Za testiranje ove hipoteze koristi se statistik $B = -[n - \frac{2m+11}{6}] \sum_{j=1}^{m} \ln \lambda_p$, pri čemu je n veličina uzorka, m broj varijabli, a λ_p svojstvene vrednosti uzoračke matrice interkorelacija. Ovaj statistik, ako je hipoteza tačna ima aproksimativno Hi-kvadrat distribuciju sa m(m-1)/2 stepeni slobode.

[109] Budući da se analiza glavnih komponenti zasniva na nalaženju svojstvenih vrednosti matrice kovarijansi ili matrice korelacija, osetljivost metode na "metriku" varijabli posledica je činjenice da svojstvene vrednosti matrice kovarijansi i matrice korelacija dobijenih od istih varijabli ne samo da nisu iste nego nisu ni u kakvim jednostavnim relacijama. Podrobnije o ovome može se pročitati u Wilks (1962).

Opisom slučajnog uzorka, tj. skupa E u pogledu skupa kvantitativnih varijabli V, pri čemu su podaci na varijablama transformisani u standardizovani oblik dobijamo matricu **Z**:

$E \otimes V = \mathbf{Z} = (z_{ij})$; i = 1,...,n; j = 1,..., m.

Matrica $\mathbf{R} = \mathbf{Z}^t\mathbf{Z}n^{-1}$ je matrica interkorelacija varijabli iz skupa V.

Formalno definisanje i rešenje problema glavnih komponenti:

Metodu glavnih komponeti je predložio H. Hotelling 1933. godine. Ova metoda se svodi se na traženje vektora koji sadrže koeficijente za linearne transformacije varijabli iz **Z** tako da na taj način dobijene linearne kombinacije (glavne komponente) merenih varijabli imaju što veću varijansu. Rečeno geometrijskim terminima, rezultujući vektor, tj. glavna komponenta treba da bude što duži, uz odgovarajuća ograničenja. Da bi formalna postavka i rešenje problema koji su dati u nastavku teksta bili lakši za praćenje zamislimo sledeću situaciju: Zadali smo uzorku ispitanika upitnik depresivnosti sa 10 pitanja. Postavlja se pitanje da li za dalju analizu moramo koristiti rezultate ispitanika na svih deset pitanja ili (najčešće, standardizovane) rezultate na pojedinačnim stavkama, tj. izvornim varijablama možemo kondenzovati (sažeti) u manji broj varijabli, glavnih komponenti, ali tako da zadržimo što veću količinu informacija koja je sadržana u odgovorima na svih 10 stavki. Ukoliko sve ove stavke mere npr. depresivnost, mogli bismo očekivati da je to moguće. Pri tom treba odlučiti kakvu ćemo težinu dati svakoj stavki (tj. odrediti ponder za linearnu kombinaciju stavki). Matematički posmatrano, na osnovu 10 stavki, tj. merenih varijabli možemo dobiti 10 glavnih komponenti. Ali neće sve ove komponente biti praktično važne i može nam se desiti ponekad da u ovakvim slučajevima zapravo možemo da umesto 10 izvornih varijabli kojima merimo depresivnost dođemo do **jedne** glavne komponente, tj. generalne mere depresivnosti. Ova, u idealnom slučaju **jedna,** glavna komponenta nastaje optimalnim ponderisanjem rezultata na pojedinačnim stavkama tako da sadrži praktično sve relevantne informacije o depresivnosti koje su sadržane u svim stavkama. Na taj način postižemo veliku uštedu (parsimoničnost) uz mali gubitak informacija.
Formalna postavka problema glavnih komponenti izgleda ovako:

$$\mathbf{Z}\,\mathbf{x}_p = \mathbf{k}_p \quad \left| \begin{array}{l} 1.\ \sigma_p^2 = \mathbf{k}_p^t\mathbf{k}_p = \mathbf{x}_p^t\mathbf{R}\,\mathbf{x}_p = \text{maximum} \\ 2.\ \mathbf{x}_p^t\mathbf{x}_p = 1 \end{array} \right.$$

Dakle, tražimo vektor $\mathbf{x}_p$ u kojem su linearni koeficijenti uz pomoć kojih iz matrice podataka **Z** dobijamo vektor, tj. glavnu komponentu $\mathbf{k}_p$ koja ima maksimalnu varijansu (uslov 1.), što znači da smo iz matrice **Z** izvukli što je moguće više informacija[110]. Naravno, ta bi varijansa bila utoliko veća ukoliko su proizvoljno veći koeficijenti u vektoru $\mathbf{x}_p$ i, stoga, uvodimo još jedan uslov (uslov 2.), a to je da je $\mathbf{x}_p$ jedinične dužine, tj. da mu je suma kvadriranih elemenata, tj. suma kvadriranih pondera za pravljenje date linearne kombinacije (glavne komponente) jednaka jedinici.
Problem rešavamo metodom Lagranžovih multiplikatora, konstruišući funkciju

$$F(\mathbf{x}_p, \lambda_p) = \mathbf{x}_p^t\mathbf{R}\,\mathbf{x}_p - \lambda_p(\mathbf{x}_p^t\mathbf{x}_p - 1)$$

[110] Varijansa je, kao što je poznato, kod kvantitativnih varijabli mera količine informacija koju varijabla sadrži.

Parcijalnom derivacijom ove funkcije po $\mathbf{x}_p$, izjednačavanjem sa 0 i skraćivanjem sa 2 dobija se jednačina

$$(\mathbf{R} - \lambda_p \mathbf{I})\, \mathbf{x}_p = \mathbf{0}$$

pa se problem tako svodi na problem svojstvenih vrednosti[111] matrice **R**, tj. matrice interkorelacija varijabli iz V.[112]

Svojstvene vrednosti λ_p matrice **R** jednake su varijansama glavnih komponenti σ_p^2. Iz $(\mathbf{R} - \lambda_p \mathbf{I})\, \mathbf{x}_p = \mathbf{0}$ sledi da je

$$\mathbf{R}\, \mathbf{x}_p - \lambda_p\, \mathbf{x}_p = \mathbf{0}$$

Ako ovu jednačinu sa leva pomnožimo sa $\mathbf{x}_p^t$ i sredimo dobijamo

$$\sigma_p^2 = \mathbf{x}_p^t\, \mathbf{R}\, \mathbf{x}_p = \lambda_p\, \mathbf{x}_p^t\, \mathbf{x}_p = \lambda_p$$

jer je

$$\mathbf{x}_p^t\, \mathbf{x}_p = 1,$$

a

$$\mathbf{x}_p^t\, \mathbf{R}\, \mathbf{x}_p = \sigma_p^2.$$

Rešenje matrične jednačine $(\mathbf{R} - \lambda_p \mathbf{I})\, \mathbf{x}_p = \mathbf{0}$ se ne može dobiti njenim množenjem sa $(\mathbf{R}-\lambda_p \mathbf{I})^{-1}$, jer kada bismo to uradili dobili bismo

$$\mathbf{x}_p = (\mathbf{R} - \lambda_p \mathbf{I})^{-1}\, \mathbf{0} = \mathbf{0}$$

što je u kontradikciji sa zahtevom da je $\mathbf{x}_p^t \mathbf{x}_p = 1$. Zato $(\mathbf{R}-\lambda_p \mathbf{I})$ ne sme imati inverz što znači da tražimo λ_p takvo da je $|\mathbf{R}-\lambda_p \mathbf{I}| = 0$, a time se rešavanje svodi na rešavanje sistema m homogenih jednačina sa m nepoznatih. Razvojem determinante dobijamo polinomsku jednačinu po λ stepena m:

$$f(\lambda) = (-\lambda)^m + b_{m-1}\, (-\lambda)^{m-1} + \ldots + b_1\, (-\lambda) + b_0 = 0$$

Ova jednačina je **karakteristična jednačina** matrice **R**, a $\lambda_1, \lambda_2, \ldots, \lambda_m$ su karakteristični koreni ili **svojstvene vrednosti** (engl. eigenvalues) matrice **R**. Njihovim nalaženjem dobijamo i njima odgovarajuće karakteristične, svojstvene vektore $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$.

Konstruišimo sada dijagonalnu matricu

$$\Lambda = (\lambda_p) \quad (p = 1, 2, \ldots, m)$$

dakle matricu čiji su dijagonalni elementi svojstvene vrednosti, poređane po veličini od gornjeg levog ka donjem desnom uglu matrice, i kvadratnu matricu

$$\mathbf{X} = (\mathbf{x}_p) \quad (p = 1, 2, \ldots, m)$$

u čijim kolonama su svojstveni vektori.

---

[111] Realni broj (skalar) λ je **svojstvena vrednost** (engl. eigenvalue, čita se kao ajgenvelju) kvadratne matrice **A** ako postoji vektor **x** sa istim brojem redova koliko ima matrica **A**, takav da mu svi elementi nisu jednaki nuli (tzv. nenulti vektor) i takav da važi $\mathbf{Ax} = \lambda \mathbf{x}$. Ukoliko izraz $\lambda \mathbf{x}$ prebacimo na levu stranu jednakosti dobijamo $\mathbf{Ax} - \lambda \mathbf{x} = 0$, tj. $(\mathbf{A} - \lambda \mathbf{I})\, \mathbf{x}$, pri čemu je **I** matrica identiteta. Dakle, svojstvene vrednosti, kada se oduzmu od dijagonalnih elemenata matrice **A** pretvaraju tu matricu u singularnu matricu, tj. matricu $\mathbf{A} - \lambda \mathbf{I}$ koja nema regularni inverz, tj. čija je determinanta $|\mathbf{A} - \lambda \mathbf{I}|$ jednaka nuli.

[112] Naravno, ukoliko se radi sa nestandardizovanim varijablama problem se svodi na rešavanje problema svojstvenih vrednosti matrice **kovarijansi**.

Budući da je matrica **R** pozitivno definitna (tj. $\mathbf{x}^t\mathbf{R}\mathbf{x} > 0$) simetrična matrica slede sledeća tvrđenja:

- $\lambda_p > 0 \quad (p = 1,...,m)$, tj. sve svojstvene vrednosti su veće od nule.
- $\det(\Lambda) = \det(\mathbf{R})$, ili, budući da je Λ dijagonalna: $\prod_{p=1}^{m} \lambda_p = \det(\mathbf{R})$
- $\text{trag}(\Lambda) = \text{trag}(\mathbf{R})$, tj. zbir svojstvenih vrednosti ili varijansi glavnih komponenti jednak je tragu matrice (zbiru vrednosti na glavnoj dijagonali matrice) iz koje su dobijene glavne komponente. Ako su glavne komponente dobijene na osnovu standardizovanih varijablii, tj. na osnovu matrice interkorelacija onda je zbir varijansi svih glavnih komponenti, kao i zbir varijansi početnih varijabli jednak broju tih varijabli.
- $\mathbf{X}^t\mathbf{X} = \mathbf{X}\mathbf{X}^t = \mathbf{I}$, tj. matrica svojstvenih vektora je ortogonalna matrica.
- $\mathbf{X}^t\mathbf{R}\mathbf{X} = \Lambda$, a odatle $\mathbf{X}\Lambda\mathbf{X}^t = \mathbf{R}$
- $\mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} = \Lambda^{-1}$, pa je $\mathbf{X}\Lambda^{-1}\mathbf{X}^t = \mathbf{R}^{-1}$, ili, u opštem slučaju $\mathbf{X}^t\mathbf{R}^w\mathbf{X} = \Lambda^w$, tj. $\mathbf{X}\Lambda^w\mathbf{X}^t = \mathbf{R}^w$ $(\forall w \in P)$. Dakle, svojstveni vektori ostaju isti za sve stepenovane matrice, pri čemu stepen pripada skupu realnih brojeva, samo se svojstvene vrednosti dižu na isti stepen. Na taj način dobijeni svojstveni vektori čine skup ortonormiranih vektora jer je $\mathbf{x}_p^t\,\mathbf{x}_q = \delta_{pq}$ pri čemu je δ_{pq} Kronekerova delta.[113] Tako dobijeni svojstveni vektori su zapravo vektori koeficijenata za glavne komponente, pri čemu je svaka glavna komponenta linearna kombinacija merenih varijabli.[114]

Matrica nestandardizovanih glavnih komponenti je

$$\mathbf{K} = \mathbf{Z}\mathbf{X}$$

a matrica kovarijansi glavnih komponenti

$$\mathbf{K}^t\mathbf{K} = \mathbf{X}^t\mathbf{Z}^t\mathbf{Z}\mathbf{X} = \mathbf{X}^t\mathbf{R}\mathbf{X} = \Lambda.$$

Očigledno je, dakle, da različite glavne komponente nisu međusobno u korelaciji[115], tj. da su vektori $\mathbf{k}_p$ ortogonalni na sve ostale vektore $\mathbf{k}_q$.

Standardizovane glavne komponente dobijamo u matrici.

$$\mathbf{K}_s = \mathbf{K}\Lambda^{-1/2} = \mathbf{Z}\mathbf{X}\Lambda^{-1/2}$$

Korelacije originalnih varijabli i glavnih komponenti dobijamo iz

$$\mathbf{F} = \mathbf{Z}^t\mathbf{K}_s = \mathbf{R}\mathbf{X}\Lambda^{-1/2} = \mathbf{X}\Lambda\Lambda^{-1/2} = \mathbf{X}\Lambda^{1/2}$$

---

[113] Ako su $\mathbf{x}_p$ i $\mathbf{x}_q$ različiti svojstveni vektori matrice **R**, onda važi i $\mathbf{R}\,\mathbf{x}_p = \lambda_p\mathbf{x}_p$ i $\mathbf{R}\,\mathbf{x}_q = \lambda_q\mathbf{x}_q$. Odatle je i $\mathbf{x}_p^t\,\mathbf{R}\,\mathbf{x}_q = \lambda_p\,\mathbf{x}_p^t\,\mathbf{x}_{pq}$ i $\mathbf{x}_q^t\,\mathbf{R}\,\mathbf{x}_p = \lambda_q\,\mathbf{x}_q^t\,\mathbf{x}_p$. Budući da je $\mathbf{x}_p^t\,\mathbf{R}\,\mathbf{x}_q - \mathbf{x}_q^t\,\mathbf{R}\,\mathbf{x}_p = 0$, sledi $(\lambda_p - \lambda_q)\,\mathbf{x}_p^t\mathbf{x}_q = 0$. Prema tome, ako $\lambda_p \neq \lambda_q$ onda $\mathbf{x}_p^t\mathbf{x}_q$ mora biti jednako 0.

[114] Geometriski posmatrano, glavne komponente su ose hiperelipsoida kojeg pravi raspored entiteta, tj. jedinica posmatranja u m - dimenzionalnom prostoru, pri čemu su koordinatne ose tog prostora izvorne, merene varijable.

[115] Glavne komponente su dakle i geometrijski i statistički ortogonalne. Geometrijska ortogonalnost postoji jer su svojstveni vektori koji odgovaraju različitim svojstvenim vrednostima međusobno ortogonalni, a statistička ortogonalnost se vidi iz toga što su kovarijanse (ili korelacije) različitih glavnih komponenti jednake nuli.

Dakle, iz $\mathbf{F} = \mathbf{Z}^t\mathbf{K}_s = \mathbf{X}\Lambda^{1/2}$, vidimo da se korelacija između određene p-te glavne komponente i izvorne j-te standardizovane varijable koja ulazi u linearnu kombinaciju za pravljenje glavne komponente dobija tako što se koeficijent za j-tu varijablu koji služe za pravljenje glavne komponente pomnoži kvadratnim korenom iz varijanse glavne komponente: $r_{jp} = x_{jp}\sqrt{\lambda_p}$.
Na osnovu toga što je

$$\mathbf{R} = \mathbf{X}\Lambda\mathbf{X}^t = \sum_{j=1}^{m} \lambda_j \mathbf{x}_j \mathbf{x}_j^t$$

vidimo da smo matricu interkorelacija raslojili na *m* nezavisnih slojeva od kojih je svaki sledeći sve "tanji", pošto su varijanse glavnih komponenti, tj. svojstvene vrednosti matrice sve manje i manje, tako da svaki sledeći sloj sve manje doprinosi rekonstrukciji matrice interkorelacija originalnih varijabli[116].

Određivanje broja važnih glavnih komponenti

Za određivanje broja značajnih, tj. važnih komponenti koje ima smisla zadržati od svih m dobijenih komponenti, postoji mnoštvo kriterijuma, ali ćemo ovde navesti samo one kriterijume koji se najčešće koriste. To su: Cattelov scree-test[117], Gutman-Kajzerov kriterijum, procenat ukupne varijanse merenih varijabli koji je obuhvaćen zadržanim glavnim komponentama.

Cattelov scree test se zasniva na grafičkoj proceni broja važnih komponenti[118]. U tu svrhu se konstruiše scater dijagram svojstvenih vrednosti, koje su, kao što smo rekli prilikom izvođenja, sortirane u nerastućem nizu. Na apcisu se nanosi redni broj svojstvene vrednosti, a na ordinatu njena vrednost. Broj komponenti se određuje tako što se iz poslednje, najmanje svojstvene vrednosti povuče linija tako da prati rast nekoliko zadnjih svojstvenih vrednosti (Grafik 18.1). Redni broj svojstvene vrednosti koja prva bude značajno preko ove prave predstavlja i broj značajnih komponenti (na grafiku 18.1 to je broj 5). U osnovi ovog postupka leži pretpostavka da su nevažne komponente posledica greške merenja i da je njihov doprinos objašnjenju varijanse skupa analiziranih varijabli zato slučajan, te se zbog toga i njima odgovarajuće svojstvene vrednosti ne razlikuju bitno jedna od druge, dakle fluktuiraju oko neke srednje vrednosti. Prva komponenta čija svojstvena vrednost odstupa od ovog trenda smatra se značajnom, tj. smatra se da ona objašnjava varijansu koja nije posledica samo greške, te iz tog razloga i odstupa od narednih svojstvenih vrednosti.

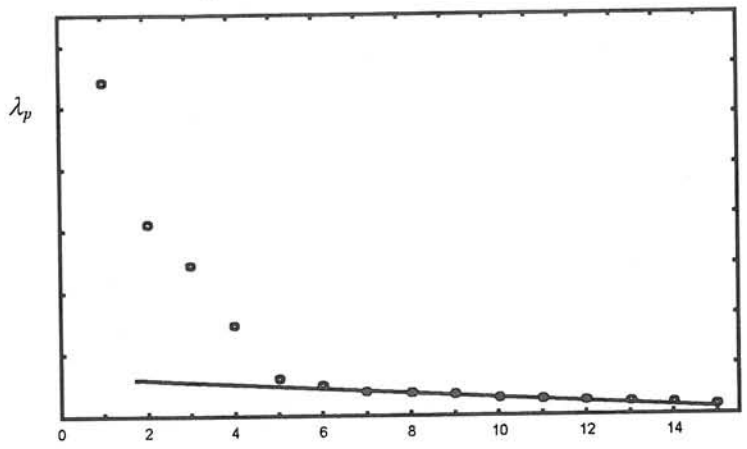
---

[116] Budući da je **R** simetrična matrica ovde je reč o posebnom slučaju teoreme o "spektralnoj dekompoziciji matrice" koja glasi: Ako je **A** simetrična matrica, postoji matrica **X** takva da je $\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^t$, pri čemu su λ_p, p = 1, ... m svojstvene vrednosti matrice **A**, a kolone matrice **X** su svojstveni vektori matrice **A**. Ime ove teoreme potiče od toga što se skup svih različitih svojstvenih vrednosti matrice zove spektralnim skupom.

[117] O testovima statističke značajnosti pri odabiranju broja glavnih komponenti koje treba zadržati neće biti govora u ovom tekstu jer se ovi testovi praktično nikad ne koriste. Distribucije uzorkovanja svojstvenih vrednosti inače su veoma kompikovane. O nekima od tih testova značajnosti može se pročitati u knjizi Z. Kovačića Multivarijaciona analiza, 1994, Beograd:Ekonomski fakultet (str. 193-195) ili S.S. Wilks (1962), Mathematical Statistics, New York: Willey & Sons, Inc. (str. 568 -573).

[118] Scree je osulina, a u geologiji se ovaj termin koristi za ruševine koje se skupljaju na donjem delu stenovitog nagiba. U ovom slučaju "ruševine" koje se skupljaju na dnu strmine su svojstvene vrednosti koje se nalaze ispod "prelomne" tačke.

Grafik 18.1: Demonstracija Cattelovog scree testa za određivanje značajnih komponenti (grafik preuzet iz diplomskog rada: A. Zorić, Faktorska analiza binarnih varijabli u generalizovanom imaž prostoru, Diplomski rad, Beograd:Filozofski fakultet, 1996.)



Guttman-Kaiser-ov kriterijum sa zasniva na meri pouzdanosti α glavne komponente koju su ovi autori konstruisali[119]:

$$\alpha_p = \frac{m}{m-1}\left(1-\frac{1}{\lambda_p}\right)$$

Kako pouzdanost može da se kreće u segmentu [0,1] to se neznačajnim komponentama proglašavaju sve one kojima je pouzdanost, po njihovom određenju mere pouzdanosti, manja od nule. Ova anomalija mere pouzdanosti se objašnjava nerealnošću date komponente, tj. ona se proglašava artificijelnom, te zato i nevažnom. Prema ovom kriterijumu treba zadržati sve komponente čija je pouzdanost veća od nule

$$k = \text{num}(\alpha_p > 0)$$

Očigledno je da je to ekvivalentno kriterijumu zadržavanja svih komponenti čija je varijansa, dakle svojstvena vrednost veća od 1 ako su izvorne varijable standardizovane:

$$k = \text{num}(\lambda_p > 1)$$

Istovremeno, glavne komponente koje imaju varijansu manju od 1 sadrže u sebi manje informacija no jedna standardizovana izvorna varijabla.

Ako se glavne komponente dobijaju na osnovu matrice kovarijansi, tj. iz nestandardizovanih izvornih varijabli, tada se zadržavaju glavne komponente sa svojstvenim vrednostima većim od proseka svih svojstvenih vrednosti[120]:

$$k = \text{num}\left(\lambda_p > \frac{\sum_{1}^{m}\lambda_p}{m}\right)$$

---

[119] Pouzdanost pokazuje stepen tačnosti s kojom se meri ono što je predmet merenja.

[120] Ponekada se umesto aritmetičke sredine u ove svrhe koristi geometrijska sredina svojstvenih vrednosti.

Kriterijum koji počiva na **procentu ukupne varijanse izvornih varijabli koja je obuhvaćena zadržanim komponentama** svodi se na to da se zadrži k glavnih komponenti tako da

$$\frac{\sum_{1}^{k} \lambda_p}{\sum_{1}^{m} \lambda_p} > p,$$

pri čemu p predstavlja proporciju ukupne varijanse merenih varijabli koja je obuhvaćena sa prvih k komponenti i određuje se najčešće kao 0.6, 0.7 ili 0.8. Dakle, treba zadržani onoliko glavnih komponenti dok proporcija ukupne varijanse izvornih varijabli koja je sadržana u tim komponentama ne pređe 60%, 70% ili 80%. Očigledno vrednost p u ovom slučaju određuje sam istraživač.

Komponentni skorovi

Svojstveni vektori matrice **R** sadrže pondere kojima se množe rezultati na pojedinim izvornim varijablama da bi se dobile glavne komponente.[121] Prema tome, matrica sa komponentnim skorovima **K** se dobija množenjem matrice podataka u kojoj su standardizovani rezultati na izvornim varijablama, matricom čije kolone sadrže svojstvene vektore matrice **R**:

$$\mathbf{K} = \mathbf{ZX}$$

Tako dobijene nekorelisane glavne komponente mogu se koristiti u daljim analizama, npr. kao prediktorske varijable u multiploj regresionoj analaizi.

Glavne komponente dobijene na osnovu standardizovanih varijabli nužno imaju aritmetičke sredine jednake nuli ali im varijanse nisu jednake jedinici.[122] Ukoliko želimo da ih standardizujemo potrebno je da svaku glavnu komponentu podelimo korenom iz njene varijanse. Matrica standardizovanih glavnih komponenti $\mathbf{K}_s$ se iz matrice **K** nestandardizovanih komponenti koje su nastale linearnom kombinacijom standardizovanih merenih varijabli dobija množenjem sa dijagonalnom matricom $\Lambda^{-1/2}$ u čijoj dijagonali su recipročne vrednosti korena varijansi glavnih komponenti:

$$\mathbf{K}_s = \mathbf{K}\Lambda^{-1/2}$$

Uslovi za korišćenje analize glavnih komponenti:

Da bi se moglo imati poverenje u rezultate dobijene analizom glavnih komponenti poželjno je da uzorak ispitanika bude ne samo slučajan već i dovoljno veliki (preko 100 ispitanika ili bar pet puta više ispitanika od broja merenih varijabli). Osim toga, budući da je u ovom

---

[121] Rezultat i-tog ispitanika na glavnoj komponentii je linearna kombinacija njegovih rezultata u standardizovanom obliku na m izvornih, merenih varijabli:

$$k_{1i} = x_{11}z_{1i} + x_{12}z_{2i} + \cdots + x_{1m}z_{mi} = \mathbf{Zx}_1$$
$$k_{2i} = x_{21}z_{1i} + x_{22}z_{2i} + \cdots + x_{2m}z_{mi} = \mathbf{Zx}_2$$
$$\cdots\cdots\cdots\cdots$$
$$k_{mi} = x_{m1}z_{1i} + x_{m2}z_{2i} + \cdots + x_{mm}z_{mi} = \mathbf{Zx}_m$$

ili u matričnom obliku: **K** = **ZX**. U matrici **K** (koja je reda n x m) su komponentni skorovi za n ispitanika na m komponenata, u matrica **Z** (reda n x m) su standardizovani rezultati n ispitanika na m izvornih varijabli, a p-ta kolona matrice **X** (reda m x m) sadrži koeficijente ili pondere za p-tu glavnu komponentu.

[122] O aritmetičkoj sredini i varijansi linearne kombinacije može se pogledati u odeljku **XIV.1.** Priručnika.

tekstu prikazan postupak koji se zasniva na matrici kovarijansi ili linearnih korelacija, potrebno je da budu ispunjeni uslovi za korišćenje koeficijenta linearne korelacije da bi se primenjivao prikazani postupak analize glavnih komponeti: podaci moraju poticati bar sa intervalne skale, distribucije varijabli treba da budu normalne (ili bar simetrične), a povezanost među varijablam linearna.

****

U SPSS programu analiza glavnih komponenti se može izvesti preko menija **Statistics/Data Reduction/Factor**. U prozoru za dijalog **Factor Analysis** merene varijable na kojima će se primeniti analiza treba prebaciti iz levog u desni okvir. Klikom na dugme **Descriptives** može se ući u poddijalog **Factor Analysis: Descriptives** u kojem je moguće, ukoliko je to potrebno zatražiti aritmetičke sredine, standardne devijacije i broj rezultata za svaku varijablu (opcija **Univariate Descriptives**), matricu interkorelacija izvornih varijabli (opcija **Coefficients** u okviru **Correlation Matrix**) ili test sferičnosti matrice interkorelacija (opcija **KMO and Bartlett's test of sphericity**). Bartletovim testom sferičnosti testira se nulta hipoteza da je populaciona matrica interkorelacija **P** jednaka matrici identiteta (H_0: **P** = **I**), tj. da li uopšte ima smisla primenjivati analizu glavnih komponenti na dobijenim podacima. Naime, ako se ova nulta hipoteza ne odbaci, analiza glavnih komponenti nema smisla.

Klikom na dugme **Extraction** u polju **Method**, treba samo proveriti da li je aktivna opcija **Principal Components**. Ova opcija se podrazumeva i ne treba je menjati za analizu glavnih komponenti. Promena polja **Method:** ima smisla samo u faktorskoj analizi. U novijim verzijama programa (SPSS 7.5 ili 8.0) moguće je (u okviru **Analyse**) definisati da li želimo analizu glavnih komponenti na osnovu standardizovanih (opcija **Correlation matrix**) ili na osnovu nestandardizovanih izvornih varijabli (opcija **Covariance Matrix**). Analiza glavnih komponenti, kao što smo već ranije istakli, nije metrički invarijantna, i, ako se radi na osnovu matrice kovarijansi, uzima više u obzir varijable koje imaju veću varijansu, a potcenjuje one sa manjom budući da je njen cilj da napravi ortogonalne komponente sa što većom varijansom. Mada ima situacija kada je analiza glavnih komponenti na osnovu matrice kovarijansi smisaona i opravdana, s obzirom na prirodu psiholoških varijabli, to najčešće nije smisaono. Stoga je, ako ne postoji jasan razlog protiv toga, najbolje ovu analizu uraditi na osnovu korelacione matrice izvornih varijabli (dakle, sa uključenom opcijom **Correlation Matrix** u novijim verzijama programa). Starije verzije programa (npr. SPSS 6.1) podrazumevaju da se analiza glavnih komponenti radi na osnovu korelacione matrice.

U okviru **Extract** treba uključiti opciju **Number of factors** i u ulazno polje treba upisati broj koji odgovara broju početnih, merenih varijabli.

U okviru **Diplay**, pored uključene opcije **Unrotated factor solution**, može se uključiti opcija **Scree plot** na osnovu kojega je moguće primeniti Katelov postupak za određivanje broja važnih komponenti.

Klikom na dugme **Scores** u glavnom prozoru faktorske analize ulazimo u poddijalog **Factor Analysis: Factor Scores** u kojem se može, uključivanjem opcije **Save as variables** zatražiti da se rezultati na glavnim komponentama snime kao poslednje varijable u radni prozor sa podacima. Uključenu opciju **Regression** u okviru **Method** ne treba menjati.

Posle klika na dugme **Options** u glavnom prozoru možemo definisati postupak u slučaju nedostajućeg podatka: **Exclude cases listwise** - izbacivanje celog ispitanika iz analize ako u samo jednoj varijabli ima nedostajući podatak, **Exclude cases pairwise** - izbacivanje ispitanika samo iz računanja korelacija za one varijable na kojima nema podatak. Opcija **listwise** je u ovoj analizi smislena, dok opcija **pairwise** nije uopšte smislena, jer se tako na

dobija prava matrica korelacija, dakle Gramova matrica[123]. Uključivanjem opcije **Replace with mean** moguće je definisati da se nedostajući podatak u nekoj varijabli zamenjuje sa aritmetičkom sredinom za datu varijablu. Postoje i dve opcije koje kontrolišu izgled matrice korelacija izvornih varijabli i glavnih komponenti: opcija **Sorted by size** omogućuje da se koeficijenti korelacije izvornih varijabli i komponenti prikažu u ispisu poređani po veličini, a opcija **Suppress absolute values less then** *r* dovodi do toga da se koeficijenti korelacija izvornih varijaili i komponenti koji su po apsolutnoj vrednosti manji od *r* ne pojavljuju u ispisu.

Primer:

Na uzorku izbeglica sa ratnih područja BiH i Hrvatske koje su pribežište našli u Srbiji (n = 198), a koji su doživele neki od ozbiljnijih traumatskih događaja tokom rata, dobijeni su odgovori na pojedinačne stavke upitnika IES (Impact of event scale-Skale uticaja događaja). Autori upitnika su Horowitz, Wilner i Alvarez a namenjen je merenju postraumatskih stresnih reakcija. Skala je bila prvobitno namenjena utvrđivanju psiholoskih posledica civilne traume a potom (pa i u ovom istrazivanju) se pokazala korisnom u ispitivanju psiholoskih posledica ratnih trauma.
Na pojedine stavke IES upitnika (npr. Mislio sam o tome i kada nisam nameravao; Pokušao sam da ne razgovaram o tome; Sve što me podseća na to ponovo me uznemiri) odgovara se zaokruživanjem jednog od ponuđenih odgovora:
Uopšte ne (0 pena);Veoma retko (1 poen);Ponekad (3 poena);Često (5 poena).
Pitanje na koje želimo da odgovorimo je da li se skup ovih 15 merenih varijabli može svesti na manji broj dimenzija koje bi u sebi sadržavale bitne informacije sadržane u izvornim varijablama. Drugim rečima da li se ovaj skup manifestnih postraumatskih reakcija može svesti na manji broj dimenzija?
Da bismo napravili analizu glavnih komponenti ovog skupa merenih varijabli odabrali smo **Statistics/Data Reduction/Factor**, a zatim u okvir **Variables** ubacili imena varijabli od IES1 do IES15. Posle klika na dugme **Descriptives**, u okviru **Correlation Matrix** uključili smo opciju **KMO and Bartlett's test of sphericity**. U poddijalogu **Extraction** ostavili smo default opciju **Principal Components**, u okviru **Extract** uključili opciju **Number of factors** i u ulazno polje upisali **15**. U okviru **Display** ostavili smo uključenu opciju **Unrotated factor solution**. Po kliku na dugme **Scores** uključena je opcija **Save as variables**, a u okviru **Method**, ostavljena je (automatski uključena) opcija **Regression**. Na taj način zatraženo je da program u aktivnom fajlu sa podacima upiše u novim varijablama (čija imena program prikazuje na kraju ispisa) komponentne skorove ispitanika. U okviru **Coefficient display format** poddijaloga **Options**, uključivanjem **Sorted by size** definisano je da se u ispisu koeficijenti korelacija izvornih varijabli i glavnih komponenti prikazuju sortirani po veličini i da se ne prikazuju korelacije koje su manje po apsolutnoj vrednosti od 0.20 (uključena je opcija **Suppress absolute values less than** i u ulazno polje upisano **.20**).

[123] Gramova ili Gramijanska matrica je simetrična i pozitivno semidefinitna ($\mathbf{x}^t\mathbf{Rx} \geq 0$). Ova matrica nastaje kao proizvod određene matrice i njenog transpona ($\mathbf{R} = \mathbf{Z}^t\mathbf{Z}$).

Ispis iz analize glavnih komponenti iz SPSS 6.1[124]:

(objašnjenja pojedinih delova ispisa su uokvirena crticama)

---------- FACTOR ANALYSIS ----------

Analysis number 1 Listwise deletion of cases with missing values

Kaiser-Meyer-Olkin Measure of Sampling Adequacy = .92892

KMO je mera reprezentativnosti merenih varijabli zą univerzum varijabli koje se odnose na dato područje. Budući da je viša od 0.80 to govori o zadovoljavajućoj reprezentativnosti.[125]

Bartlett Test of Sphericity = 1715.2642, Significance = .00000

Budući da je Significance manje od .05 odbacujemo nultu hipotezu da je matrica interkorelacija merenih varijabli u populaciji jednaka matrici identiteta., što znači da ima smisla raditi analizu glavnih komponenti na datim varijablama.

Extraction 1 for analysis 1, Principal Components Analysis (PC)
Initial Statistics:

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
IES1	1.00000	*	1	7.31400	48.8	48.8
IES2	1.00000	*	2	1.53233	10.2	59.0
IES3	1.00000	*	3	1.04987	7.0	66.0
IES4	1.00000	*	4	.79555	5.3	71.3
IES5	1.00000	*	5	.70775	4.7	76.0
IES6	1.00000	*	6	.66197	4.4	80.4
IES7	1.00000	*	7	.49282	3.3	83.7
IES8	1.00000	*	8	.41794	2.8	86.5
IES9	1.00000	*	9	.37669	2.5	89.0
IES10	1.00000	*	10	.36448	2.4	91.4
IES11	1.00000	*	11	.32272	2.2	93.6
IES12	1.00000	*	12	.30965	2.1	95.6
IES13	1.00000	*	13	.24115	1.6	97.2
IES14	1.00000	*	14	.22127	1.5	98.7
IES15	1.00000	*	15	.19179	1.3	100.0

PC extracted 15 factors.

U koloni **Eigenvalue** su svojstvene vrednosti matrice interkorelacija a u koloni **Pct of var** procenat ukupne varijanse svih merenih varijabli koji je obuhvaćen datom glavnom komponentom. U koloni **Cum Pct** je procenat ukupne varijanse merenih varijabli koji je obuhvaćen datom komponentom i svim komponentama pre nje. Treba uočiti da je zbir svojstvenih vrednosti jednak 15, tj. broju varijabli. To je uvek slučaj kada glavne

---

[124] Budući da je ispis iz ove metode praktično isti i u novijim verzijama programa ovde će biti dat ispis samo iz verzije 6.1, dok je prikazani grafik Scree testa napravljen u verziji 8.0 ovog programa.

[125] O ovoj meri se može pročitati u K. Momirović, B. Wolf & D.A. Popović (1999). Uvod u teoriju merenja I (str. 177).

komponente ekstrahujemo iz matrice interkorelacija, jer je tada varijansa svake izvorne varijable jednaka 1.

Factor Matrix[126]:

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
IES5	.84694	-.24407			
IES14	.81931	-.21023			
IES10	.81714	-.30879			
IES7	.75377	.40631			
IES13	.74361	.44439			
IES1	.73602	-.31559	-.21449		
IES4	.73298				-.37912
IES12	.72901	-.32217			
IES6	.72549				-.33474
IES11	.68684	-.31617			.43260
IES2	.68383		-.32265	.22764	.35778
IES3	.64801	.49851	-.24949		
IES9	.62206	.46391			
IES8	.30588	.38983	.66545	-.31362	
IES15	.40001		.53152	.72762	

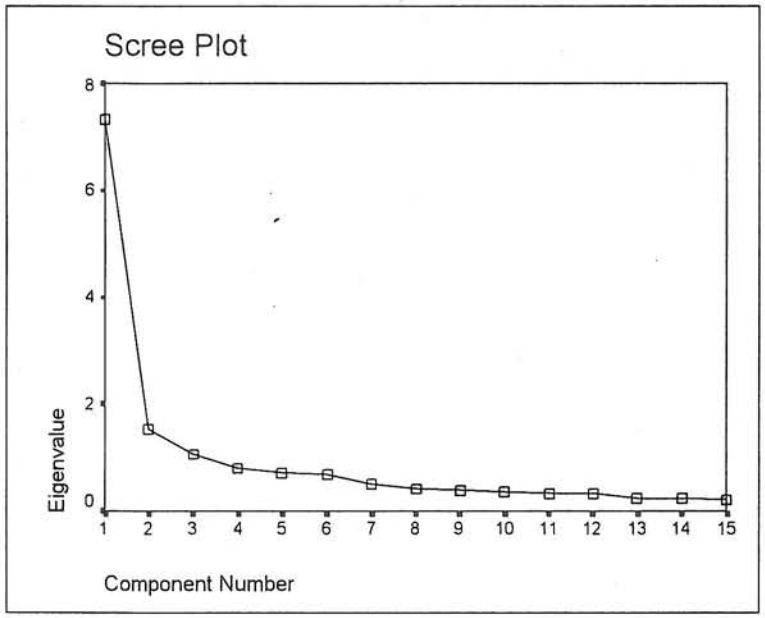
	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
IES5					
IES14			-.24975		
IES10					
IES7			-.25945		
IES13					
IES1	.26293			.32462	
IES4		.28399	.38716		
IES12		.32431			-.41093
IES6		-.43350			-.27843
IES11				-.36122	
IES2	.22952	-.25921			
IES3	.25291				
IES9	-.43047		.27344	.20220	
IES8	.42351				
IES15					

	Factor 11	Factor 12	Factor 13	Factor 14	Factor 15
IES5			-.27677	.20024	-.20638
IES14				.20656	
IES10				-.35157	
IES7	.20630				.21815
IES13	-.21175	-.30899			-.20684
IES1	.24642				
IES4					
IES12					
IES6					
IES11					
IES2					
IES3		.28813			
IES9					
IES8					
IES15					

Koeficijenti u, kako je SPSS zove, "faktorskoj" matrici (**Factor matrix**) su koeficijenti linearnih korelacija svake merene varijable i glavnih komponenti. Ovi su koeficijenti sortirani po veličini jer smo to pri definisanju procedure tražili, a prazna mesta su umesto koeficijenata koji su manji od 0.20, što smo takođe zahtevali pri definisanju procedure

[126] U verziji 8.0 programa SPSS ova se matrica zove **Component Matrix**

(uključivanjem **Sorted by size** i **Suppress absolute values less than 0.20** u okviru **Coefficient display format** u poddijalogu **Options**).



Grafik pod naslovom **Scree Plot** pokazuje promene u veličinama svojstvenih vrednosti idući od prve do poslednje glavne komponente

---

Final Statistics:

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
		*				
IES1	1.00000	*	1	7.31400	48.8	48.8
IES2	1.00000	*	2	1.53233	10.2	59.0
IES3	1.00000	*	3	1.04987	7.0	66.0
IES4	1.00000	*	4	.79555	5.3	71.3
IES5	1.00000	*	5	.70775	4.7	76.0
IES6	1.00000	*	6	.66197	4.4	80.4
IES7	1.00000	*	7	.49282	3.3	83.7
IES8	1.00000	*	8	.41794	2.8	86.5
IES9	1.00000	*	9	.37669	2.5	89.0
IES10	1.00000	*	10	.36448	2.4	91.4
IES11	1.00000	*	11	.32272	2.2	93.6
IES12	1.00000	*	12	.30965	2.1	95.6
IES13	1.00000	*	13	.24115	1.6	97.2
IES14	1.00000	*	14	.22127	1.5	98.7
IES15	1.00000	*	15	.19179	1.3	100.0

Skipping rotation 1 for extraction 1 in analysis 1

15 PC EXACT factor scores will be saved.

```
Following factor scores will be added to the working file:

  Name        Label

FAC1_1        REGR factor score   1 for analysis    1
FAC2_1        REGR factor score   2 for analysis    1
FAC3_1        REGR factor score   3 for analysis    1
FAC4_1        REGR factor score   4 for analysis    1
FAC5_1        REGR factor score   5 for analysis    1
FAC6_1        REGR factor score   6 for analysis    1
FAC7_1        REGR factor score   7 for analysis    1
FAC8_1        REGR factor score   8 for analysis    1
FAC9_1        REGR factor score   9 for analysis    1
FAC10_1       REGR factor score  10 for analysis    1
FAC11_1       REGR factor score  11 for analysis    1
FAC12_1       REGR factor score  12 for analysis    1
FAC13_1       REGR factor score  13 for analysis    1
FAC14_1       REGR factor score  14 for analysis    1
FAC15_1       REGR factor score  15 for analysis    1
```

Ispis pod naslovom **Final Statistics** potpuno je istovetan onom pod naslovom **Initial Statistics**. Razlike u ova dva dela ispisa se pojavljuju tokom rotacije faktora u faktorskoj analizi. Kolona **Communality** u oba slučaja sadrži jedinice za svaku merenu varijablu. Komunaliteti pojedinih merenih (ili u terminima faktorskog modela manifestnih) varijabli takođe nisu od interesa u analizi glavnih komponenti. Pojam komunaliteta veoma je važan u faktorskoj analizi u kojoj predstavlja proporcija ukupne varijanse svake manifestne varijable koja je objašnjena zajedničkim faktorima.

Deo ispisa na kraju ispod naslova `Following factor scores will be added to the working file:` služi kao obaveštenje o imenima varijabli (kolona **Name**) u kojima su, u aktivnom fajlu sa podacima, zapisani skorovi ispitanika na glavnim komponentama.

Pre odgovora na pitanje koje smo postavili pre izvođenja analize glavnih komponenti treba uočiti da bi prema Gutman-Kajzerovom kriterijumu postojale tri važne glavne komponente. Sličnu sugestiju daje i Katelov Scree test. Ukoliko bismo zadržali prve tri glavne komponente onda bismo njima obuhvatili 66% varijanse svih merenih varijabli, tj. svih stavki iy upitnika. Međutim, pre konačne odluke dobro je pokušati sa interpretacijom glavnih komponenti. To, naravno podrazumeva dobro poznavanje sadržaja merenih varijabli. Na osnovu matrice korelacija pojedinih stavki i glavnih komponenti (Factor matrix) vidimo da sa njom najvišu korelaciju pokazuju stavke 5 ("Navirala su mi snažna osećanja u vezi sa tim"), 14 ("Sve što me podseća na taj događaj ponovo mi vraća i osećanja vezana za njega") i 10 ("Te slike se same javljaju i kad mislim o nečem drugom"). Očigledno radi se o komponenti koja u sebi sadrži elemente nametanja, tj. nevoljnog ponovnog preživljavanja traumatskog događaja. Druga komponenta ima najviše korelacije sa stavkama 3 ("Pokušao sam da to izbacim iz sećanja"), 9 ("Pokušao sam da ne pričam o tome"), 7 ("Izbegavao sam sve što me na taj događaj podseća") i 13 ("Pokušao sam da ne mislim o tome") i predstavlja reakcije izbegavanja stimulusa u vezi sa traumom koje nisu u vezi sa nametanjem. Reakcije izbegavanja koje su u vezi sa nametanjem sadržane su u prvoj glavnoj komponenti. Treća komponenta korelira najviše sa stavkama 8 ("Imao sam osećaj kao da se to nije desilo ili nije bilo stvarno") i 15 ("Moja osećanja vezana uz taj događaj kao da su zaleđena") i podseća na reakciju poricanja (denial) i emocionalne otupljenosti.

Dakle, skup od 15 merenih varijabli (stavki) mogao bi se u ovom slučaju transformisati, tj. svesti na tri glavne komponente koje bi u sebi sadržale sve bitne informacije, tj. veliki deo

varijanse 15 merenih varijabli. Naravno, treća zadržana komponenta je u izvesnoj meri problematična jer sadrži uglavnom informacije iz dveju stavki. Detaljnija analiza upitnika mogla bi dovesti do toga da ove stavke eliminišemo iz upitnika i da se, na taj način, celokupna relevantna informacija koja je sadržana u odgovorima ispitanika na preostalih 13 stavki svede na dve glavne komponente.

****

XIX. ANALIZA TRENDA

Analiza trenda spada u širu oblast analize vremenskih serija. Analiza vremenskih serija ima veliku važnost u proučavanju psiholoških pojava. Razvojne promene u mnogim psihičkim funkcijama tokom ljudskog veka, kojima se bavi razvojna psihologija, praćenje efekata pedagoškog ili psihoterapijskog tretmana, praćenje produktivnosti određene proizvodne grupe tokom niza meseci, promene u učenju i zaboravljanju tokom vremena, promene popularnosti određenih političkih lidera, samo su neki primeri psiholoških problema u kojima analiza vremenskih serija ima svoju primenu.

XIX.1. Pojam vremenskih serija

Vremenska serija je skup rezultata posmatranja nekog procesa tako da se posmatranja vrše u sukcesivnim vremenskim trenucima. Dakle, vrednosti određenih obeležja za iste jedinice posmatranja snimane u jednakim vremenskim razmacima čine vremensku seriju.
Vremenske serije mogu biti:

1. **Neprekidne** - vrednosti vremenske serije određene su u svakom proizvoljnom vremenskom trenutku (npr. zapis elektroencefalograma)
2. **Diskretne** - vrednosti vremenske serije određene su u zadatim prekidnim vremenskim trenucima (npr. broj pacijenata koji dođu psihologu na svakih sat vremena; broj udesa po danima, broj reprodukovanih prethodno naučenih slogova na svaki sat po završenom učenju). Diskretne vremenske serije se, pak, mogu podeliti na :
   Momentne, tj. takve u kojima se vrednosti vremenske serije posmatraju u zadatim, ekvidistantnim vremenskim tačkama. U takve serije spada, na primer, broj tačno reprodukovanih reči posle svakog sata od trenutka kada su sve reči naučene.
   Intervalne, tj. serije u kojima se vrednsoti vremenske serije dobijaju usrednjavanjem pokazatelja za određeni vremenski interval. Primer takvih vremenskih serija bila bi prosečna količina potrošenih lekova za svaki mesec tokom određenog vremenskog perioda.

Vrednost vremenske serije u trenutku t (koju označavamo sa $v_{(t)}$) teorijski se sastoji iz dve komponente: determinisane ili sistematske ($v^*_{(t)}$) i slučajne komponente ili šuma ($\gamma_{(t)}$):

$$v_{(t)} = v^*_{(t)} + \gamma_{(t)}$$

Statističke metode analize vremenskih serija mogu se podeliti u dve grupe:

1. Metode koje analizi vremenskih serija prilaze iz domena frekvenci - harmonijska i spektralna analiza;
2. Metode koje analizi vremenskih serija prilaze iz domena vremena. U ove metode spadaju modeli razvojnih tendencija, metode za analizu međusobnih odnosa vremenskih serija i metode za predviđanje vremenskih serija u funkciji vremena.

Pri empirijskoj analizi promena određene pojave u funkciji vremena, varijacije vremenskih serija možemo razložiti na četiri ključne komponente:

1. Opšta razvojna tendencija pojave u funkciji vremena- trend ($t_{(t)}$);
2. Ciklična kolebanja koja se ponavljaju u (često nejednakim) razmacima ($c_{(t)}$);

3. Sezonske varijacije koje se javljaju u razmacima manjim od 1 godine i ponavljaju se u dužem nizu godina ($s_{(t)}$)
4. Rezidualne (slučajne) varijacije ($g_{(t)}$)

Prve tri komponente čine sistematsku komponentu vremenske serije iz teorisjkog modela. Postoje različite teorije o međusobnim odnosima pojedinih komponenti vremenskih serija:

I. Multiplikativna: $y_{(t)} = t_{(t)} * c_{(t)} * s_{(t)} * g_{(t)}$
II. Aditivna: $y_{(t)} = t_{(t)} + c_{(t)} + s_{(t)} + g_{(t)}$;
III. Hibridna: $y_{(t)} = t_{(t)} + (c_{(t)} * s_{(t)} * g_{(t)})$; itd.

U nastavku ovog teksta biće prikazani samo najosnovniji principi analize polinomijskih trendova, tj. analize opšte tendencije promene **jedne** varijable u funkciji vremena[127].

XIX.2. Analiza polinomskih trendova

Struktura podataka:

v je zavisna, kriterijska varijabla za koju vrednosti koje uzima snimamo u ekvidistantnim vremenskim tačkama;
$t_1, t_2, t_3, \ldots, t_f$ su ekvidinstantne vremenske tačke;
t je vektor čiji elementi su $t_1, t_2, t_3, \ldots, t_f$
y je vektor u kojem su vrednosti varijable v u vremenskim tačkama $t_1, t_2, t_3, \ldots, t_f$.

Formalno definisanje problema:

$$\mathbf{Tb} = \mathbf{y} - \mathbf{g} \mid \mathbf{g}^t\mathbf{g} = \min$$

$$\mathbf{b} = ?$$

Matrica **T** je matrica reda f x f čija prva kolona ima f jedinica a ostalih f - 1 kolona služe za određivanje koeficijenata polinoma. U preostalih f - 1 vektora su: u prvom sledećem vektoru celi brojevi od 1 do f, zatim kvadrati ovih brojeva, treći stepeni celih brojeva i tako redom. Izraz $\mathbf{g}^t\mathbf{g}$ označava sumu kvadriranih reziduala (grešaka), tj. zbir kvadriranih razlika empirijski dobijenih i očekivanih (na osnovu modela) vrednosti kriterijumske varijable v.

Problem je, dakle, naći osnovnu tendenciju vremenskog niza, složaj vremenske serije, zakonitost po kojoj se menjaju rezultati varijable v u funkciji vremena, tj. naći što jednostavniju matematičku funkciju koja što je moguće bolje opisuje zakonitost promene vrednosti varijable v u funkciji vremena. Potrebno je opisati što je moguće tačnije skup tačaka koje definišu dobijene vrednosti varijable v u vremenskim tačkama $t_1, t_2, t_3, \ldots, t_f$ polinomom što je moguće nižeg stepena, tj. što je moguće jednostavnijom matematičkom funkcijom.

---

[127] Posebna oblast analize vremenskih serija je ispitivanje povezanosti dveju ili više varijabli tako da se vrednosti tih varijabli snimaju na istim jedinicama posmatranja tokom vremena. Tako, na primer, moguće je ispitivati efekat stepena nezaposlenosti na nivo kriminala snimanjem promena obe ove pojave tokom dužeg niza godina. Budući da između sukcesivnih vrednosti na tim varijablama u ovakvom slučaju postoji tzv. autokorelisanost reziduala, tj. grešaka, nije dozvoljeno primenjivati standardni regresioni model (koji je objašnjen u odeljcima **XI** i **XIV** ovog teksta).

Opšti oblik polinoma za aproksimaciju trenda je:

$$y = \sum_{p=0}^{s} b_p t^p + g = b_o + b_1 t + b_2 t^2 + \ldots + b_s t^s + g$$

Bez obzira na to što se raspored od f tačaka može savršeno opisati polinomom stepena f-1, ne treba zaboraviti da uvođenje složenijeg modela podrazumeva dodavanje novog parametra i (zbog neophodnosti njegove ocene na osnovu podataka) gubitka jednog stepena slobode.

Rešenje na osnovu principa najmanjih kvadrata (za polinome reda manjeg od 5)

$$\mathbf{b} = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{y}$$

b je vektor koeficijenata polinoma.
Očekivane vrednosti varijable v u vremenskim tačkama t_1, t_2, t_3,..., t_f nalaze se u vektoru $\mathbf{y}^* = \mathbf{Tb} = \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{y}$. Deljenjem $\mathbf{g}^t\mathbf{g}$, tj. zbira kvadriranih reziduala sa (f -p -1) dobijamo varijansu reziduala, (p je broj parametara, tj. koeficijenata koje ocenjujemo na osnovu podataka):

$$S_g^2 = \frac{\sum_{t=1}^{f} g_t^2}{f-p-1} = \frac{\sum_{t=1}^{f} (y_t - y_t^*)^2}{f-p-1}$$

Mere tačnosti, tj podesnosti (engl. fit) modela u aproksimiranju rasporeda empirijskih tačaka su korelacija predviđenih i empirijski dobijenih vrednosti varijable v i standardna greška ocene, tj. standardna devijacija distribucije reziduala ili kvadratni koren varijanse reziduala.
Za određivanje koji vektori u **T** značajno doprinose objašnjenju trenda može poslužiti parcijalna korelacija pojedinih elemenata (komponenti) polinoma i empirijski dobijenih vrednosti varijable v.

Predviđanje vrednosti kriterijumske varijable

Za razliku od interpolacije koja se odnosi na određivanje (na osnovu modela) vrednosti zavisne varijable v u intervalu između sukcesivnih vrednosti u ispitivanom rasponu vremenskih tačaka, ekstrapolacija se odnosi na prognozu vrednosti zavisne varijable u nekom budućem vremenskom trenutku. Pri ekstrapolaciji treba biti veoma oprezan jer je ona zavisna od toga da li model važi i za buduća kretanja vrednosti varijable v.
Za polinome reda višeg od 5 koeficijenti u vektoru **b**, dobijeni metodom najmanjih kvadrata, postaju vrlo nesigurni, pa se do rešenja dolazi metodom ortogonalnih polinoma.

Rešenje problema metodom ortogonalnih polinoma (za polinome čiji red je viši od 5)

Konstruišemo matricu Ω koja ima f ortogonalnih polinoma, tj. matricu čiji svaki par vektora kada se skalarno pomnože daju nulu. Dakle $\Omega^t\Omega = \mathbf{D}$, gde je **D** dijagonalna matrica.
Rešenjem problema:

$$\Omega\alpha = \mathbf{y} - \mathbf{g} \mid \mathbf{g}^t\mathbf{g} = \min$$

dobijamo

$$\alpha = \mathbf{D}^{-1}\Omega^t\mathbf{y}.$$

Matrica ortogonalnih polinoma konstruiše se na osnovu vektora čiji koeficijenti su dati u detaljnijim udžbenicima statistike. Određivanje dodatnog doprinosa svakog narednog člana

polinoma podesnosti modela za opis trenda u ovom slučaju je jednostavniji, nego pri korišćenju metode najmanjih kvadrata, jer su članovi polinoma međusobno ortogonalni pa se ovaj dodatni doprinos svakog sledećeg člana polinoma može odrediti prostom korelacijom toga člana i kriterijumske varijable. Na taj način moguće je prosto dodavanjem jednog po jednog člana polinoma proces zaustaviti kada se dobije podesan model. Budući da matrica $\mathbf{T}^t\mathbf{T}$ koja je neophodna za dobijanje vektora **b** u prethodnom postupku ne može biti dijagonalna (budući da su vandijagonalni elementi zbir stepena iste varijable t), ceo skup koeficijenata se mora računati ponovo svaki put kada se novi član dodaje u polinom.

****

U SPSS programu odgovor na pitanje kojim polinomom najnižeg stepena (ili kojom najjednostavnijom matematičkom funkcijom) možemo dobro da aproksimiramo trend neke pojave u vremenu možemo uraditi odabirom menija **Statistics/Regression/Curve Estimation**. U dijalogu **Curve Estimation** kao **Dependent(s)** varijablu treba uneti ime zavisne, kriterijumske varijable v a u okviru **Independent** odabrati vrema kao prediktor (opcija **Time**). Ukoliko želimo grafički prikaz rezultata analize možemo uključiti opciju **Plot Models**. U okviru **Models** treba odabrati željeni model i uključiti opciju **Display Anova Table** ispod ovog okvira. Ponuđeni modeli (polinomski i drugi) u okviru **Models** su:

1. Linearni (**Linear**): $y^* = b_0 + b_1 t$
2. Kvadratni (**Quadratic**): $y^* = b_0 + b_1 t + b_2 t^2$
3. Kubični (**Cubic**): $y^* = b_0 + b_1 t + b_2 t^2 + b_3 t^3$
4. Eksponencijalni (**Exponential**): $y^* = b_0 e^{b1\,t}$
5. Modifikovani eksponencijalni (Compound): $y^* = b_0 b_1^t$
6. Model rasta (**Growth**): $y^* = e^{b0 + b1\,t}$
7. Logaritamski (**Logarithmic**): $y^* = b_0 + b_1 \ln(t)$
8. S model (**S**) : $y^* = e^{b0 + b1/t}$
9. Hiperbolički (**Inverse**) $y^* = b_0 + b_1/t$
10. Stepeni ili parabolički (**Power**): $y^* = b_0\, t^{b1}$
11. Logistički (**Logistic**): $y^* = 1 / (1/u + b_0\, b_1^t)$. U ovom modelu oznakom u obeležena je gornja granična vrednost koja mora biti pozitivna i veća od najveće vrednosti u kriterijumskoj varijabli.

Predviđena kriterijumska varijabla je u gornjim jednačinama označena sa y^*, prediktorska varijabla, tj vreme je t, dok su b_0 , b_1, b_2, i b_3 ocene parametara, tj. koeficijenti modela[128].

U opciji **Save** može se zahtevati zapisivanje u aktivnom fajlu sa podacima novih varijabli sa predviđenim (na osnovu modela) vrednostima kriterijumske varijable (**Predicted values**), reziduala (**Residuals**) i 90%, 95% ili 99% interval poverenja za predviđene vrednosti (odabirom opcije **Prediction intervals** i odabirom odgovarajućeg broja u polju ispred **% Confidence interval**). Gornje i donje granice ovog intervala u aktivnom fajlu će biti sačuvane kao posebne varijable.

U delu ispisa koji nosi naslov **Variables in the Equation** date su u koloni **B** ocene parametara (koeficijenata) matematičkog modela trenda.

Adekvatnost matematičke funkcije (modela) za opis trenda varijable v može se odrediti poređenjem vrednosti statistika **Multiple R** (korelacija predviđenih modelom i empirijskih

---

[128] U svim modelima e je oznaka osnove prirodnih logaritama, koja je, kao što je poznato jednaka približno 2.718. Inače matematički izraz za e je:

$$e = \lim_{n\to\infty}(1+\tfrac{1}{n})^n$$

vrednosti kriterijumske varijable) i **Standard Error** (standardna devijacija distribucije reziduala) za različite modele kao i pregledanjem grafika, tj. na osnovu ispitivanja koliko je kriva, definisana modelom, blizu empirijskim tačkama. Što je korelacija veća a standardna greška manja to je model adekvatniji. Opredeljujemo se za matematički složeniji model samo ako je znatno superiorniji od jednostavnijeg u opisu trenda varijable v.

Primer1:

Pokazatelj uspeha ispitanika u određenoj aktivnosti sniman je na svaki sat. Kojim modelom se najbolje može opisati trend uspeha, tj. generalna tendencija promene uspeha u funkciji vremena. U ispisu iz ove analize dati su pokazatelji podesnosti linearnog i kvadratnog modela:

```
MODEL:  MOD_6.

Dependent variable.. USPEH1              Method.. LINEAR

Listwise Deletion of Missing Data

Multiple R            .99871
R Square              .99742
Adjusted R Square     .99728
Standard Error       2.01073

            Analysis of Variance:

                DF    Sum of Squares      Mean Square

Regression       1        29648.421        29648.421
Residuals       19           76.817            4.043

F =     7333.24235       Signif F =  .0000

-------------------- Variables in the Equation --------------------

Variable                    B         SE B       Beta          T  Sig T

Time                 6.205195      .072462    .998707     85.634  .0000
(Constant)          -3.733333      .909866                -4.103  .0006


Dependent variable.. USPEH1              Method.. QUADRATI

Listwise Deletion of Missing Data

Multiple R            .99871
R Square              .99742
Adjusted R Square     .99713
Standard Error       2.06566
```

```
            Analysis of Variance:

                DF   Sum of Squares      Mean Square

Regression       2        29648.433        14824.217
Residuals       18           76.805            4.267

F =     3474.20259       Signif F =  .0000

------------------- Variables in the Equation -------------------

Variable                    B        SE B       Beta          T  Sig T

Time                 6.188850     .312416    .996076     19.810  .0000
Time**2               .000743     .013792    .002709       .054  .9576
(Constant)          -3.670677    1.492149               -2.460  .0242
```

Zaključak: Budući da je koeficijent korelacije (**Multiple R**) za linearni model viši i iznosi 0.999, a da je standardna greška (**Standard Error**)za linearni model 2.01 i nešto je niža od standardne grečke složenijeg kvadratnog modela, mogli bismo se u ovom slučaju opredeliti za linearni model USPEH1 = -3.73 + 6.21*t kao adekvatniji opis trenda uspeha od modela USPEH1 = -3.67 + 6.19*t + 0.0007*t^2

Primer2: Kojim najjednostavnijim modelom bismo mogli opisati trend promene prosečnog broja grešaka (varijabla BRGR3) tokom određenog eksperimenta?
U ispisu su dati pokazatelji podesnosti linearnog, logaritamskog, kvadratnog i kubičnog modela:

```
Dependent variable.. BRGR3                Method.. LINEAR

Listwise Deletion of Missing Data

Multiple R            .15869
R Square              .02518
Adjusted R Square    -.02612
Standard Error      58.27823

            Analysis of Variance:

                DF   Sum of Squares      Mean Square

Regression       1         1667.129        1667.1286
Residuals       19        64530.681        3396.3516

F =         .49086       Signif F =  .4920

------------------- Variables in the Equation -------------------

Variable                    B        SE B       Beta          T  Sig T

Time               -1.471429    2.100201   -.158695      -.701  .4920
(Constant)         60.947619   26.371261                 2.311  .0322

Dependent variable.. BRGR3                Method.. LOGARITH

Listwise Deletion of Missing Data
Multiple R            .45183
R Square              .20415
Adjusted R Square     .16226
Standard Error      52.65759
```

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	1	13514.195	13514.195
Residuals	19	52683.614	2772.822

F = 4.87381 Signif F = .0398

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
Time	-31.793729	14.401487	-.451828	-2.208	.0398
(Constant)	113.466849	33.174648		3.420	.0029

Dependent variable.. BRGR3 Method.. QUADRATI

Listwise Deletion of Missing Data

Multiple R	.96039
R Square	.92234
Adjusted R Square	.91371
Standard Error	16.89964

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	2	61057.051	30528.526
Residuals	18	5140.759	285.598

F = 106.89346 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
Time	-37.267767	2.555948	-4.019360	-14.581	.0000
Time**2	1.627106	.112833	3.975160	14.420	.0000
(Constant)	198.166917	12.207627		16.233	.0000

Dependent variable.. BRGR3 Method.. CUBIC

Listwise Deletion of Missing Data

Multiple R	.96264
R Square	.92667
Adjusted R Square	.91373
Standard Error	16.89817

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	3	61343.488	20447.829
Residuals	17	4854.321	285.548

F = 71.60900 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
Time	-30.895058	6.856902	-3.332058	-4.506	.0003
Time**2	.919504	.715455	2.246427	1.285	.2160
Time**3	.021442	.021409	1.075802	1.002	.3306
(Constant)	185.147034	17.832318		10.383	.0000

Zaključak: Sudeći po visini korelacije i standardnih grešaka, kvadratni i kubični model su podjednako dobri, ali se iz razloga parsimoničnosti ili jednostavnosti u ovom slučaju opredeljujemo za kvadratni model. Dakle model BRGR3 = 198.17 -37.27*t + 1.63*t^2 bi bio najpodesniji model za opis trenda prosečnog broja grešaka u eksperimentu.

****

DODATAK 1: OSNOVE RADA NA PC RAČUNARU POD OPERATIVNIM SISTEMOM WINDOWS I OSNOVE KORIŠĆENJA INTERNET EKSPLORERA

SADRŽAJ DODATKA 1:

I. UVODNI POJMOVI 3

II. DELOVI PERSONALNOG RAČUNARSKOG SISTEMA 6

III. OSNOVE RADA NA PERSONALNOM RAČUNARU SA OPERATIVNIM SISTEMOM WINDOWS 8

A. OPERATIVNI SISTEM 8

B. UKLJUČIVANJE I ISKLJUČIVANJE RAČUNARA 8

1. Uključivanje računara 8

2. Isključivanje računara 10

C. MIŠ I TASTATURA 10

1. Rukovanje mišem 10

2. Tastatura 11

D. FORMATIRANJE DISKA 14

E. DRAJV 15

F. FOLDER (DIREKTORIJUM) 15

G. FAJL 15

H. OSNOVNE OPERACIJE U WINDOWS 3.xx I WINDOWS95 OKRUŽENJU 18

1. Rad sa prozorima (windows-ima) 18

2 Aktiviranje programa u Windows okruženju 22

3. Rad sa diskovima, folderima i fajlovima u Windows okruženju 23

4. Aktiviranje programa koji rade pod DOS-om u Windows95 okruženju 27

IV. OSNOVE KORIŠĆENJA INTERNET EKSPLORERA 27

A. INTERNET 27

B. ZNAČENJE NAJVAŽNIJIH TERMINA U VEZI SA INTERNETOM 28

C. INTERNET EXPLORER 29

I. UVODNI POJMOVI

RAČUNAR

je mašina univerzalne namene jer uz pomoć odgovarajućeg PROGRAMA može poslužiti kao kalkulator, najsavršenija pisaća mašina, elektronski klavir, psihoterapeut(!)...

PROGRAMI (SOFTWARE)

se sastoje iz računaru razumljivo napisanih algoritama za rešavanje određenih problema. Sastavljeni su od niza instrukcija koje upravljaju HARDVEROM i tako univerzalnu mašinu pretvaraju u mašinu specijalne namene. Skup programa obično krstimo imenom SOFTVER (SOFTWARE).

HARDVER (HARDWARE)

je skup svih elektronskih i mehaničkih komponenti računara. Elektronika današnjih računara je elektronika elektronskih kola visokog nivoa integracije. Elektronsko kolo je u suštini bistabilni element (može biti u dva stanja). Ukoliko na nekom dovodu ima napona to je stanje logičke jedinice a ukoliko napona nema reč je o stanju logičke nule. Prema tome sve ono što možemo stvoriti uz pomoć računara u krajnjem se slučaju svodi na nepregledno mnoštvo BINARNIH stanja pojedinih elektronskih kola računara. Dešavanja unutar računara se, stoga, najbolje mogu opisati Bulovom algebrom.

INFORMACIJA

predstavlja saopštenje o događaju u nekom sistemu. U teoriji informacija pojava koja je određena ovim pojmom potpuno je oslobođena atributa materijalnosti, tj. forme u kojoj je data.

KOLIČINA INFORMACIJE

je jednaka broju potrebnih binarnih opredeljenja da bi se fiksirao neki događaj u sistemu, tj. jednaka je binarnom logaritmu broja mogućih događaja $h = \log_2 N$- pri čemu je h=količina informacije, N=broj mogućih stanja nekog sistema.

BINARNI BROJNI SISTEM

Osnovu ovog sistema čine cifre 0 i 1. Računar operiše direktno sa binarnim brojevima a ljudi obično koriste DECIMALNI brojni sistem. U računarstvu pokatkad imamo posla i sa OKTALNIM (cifre od 0 do 7) i HEKSADECIMALNIM brojnim sistemima (cifre 0 1 2 3 4 5 6 7 8 9 A B C D E F).
Svi ovi sistemi spadaju u grupu POZICIONIH brojnih sistema jer vrednost pojedine cifre zavisi od njene pozicije u broju: vrednost cifre 5 u brojevima 5986 i 25 je očigledno različita. U APOZICIONIM brojnim sistemima to nije slučaj: vrednost X u rimskim brojevima XI i LX je ista mada su im pozicije različite.

BIT

je binarni broj (**BI**NARY DIGIT) 1 ili 0 i osnovna je jedinica količine informacija. Količina informacije koja odgovara jednom bitu jednaka je jednom opredeljenju između dve jednako verovatne mogućnosti.

BAJT (BYTE)

se sastoji od 8 bita. Najčešće se jednim bajtom u računaru predstavlja jedan karakter (slovo ili cifra).

K (KILOBAJT)

se sastoji od 1024 bajta, a 1024 Kilobajta čine MEGABAJT (MEGABYTE). GIGABAJT čine 1024 megabajta a TERABAJT je 1024 gigabajta. U gornjim jedinicama izražavamo kapacitet MEMORIJE računara.

REČ (RAČUNARSKA, MAŠINSKA)

je niz bitova koje računarski hardver tretira kao jedinicu (količina memorije koju procesor može uzeti i obraditi odjednom). Reč je dovoljno duga da primi instrukciju ili ceo broj. Najčešće ima 16 i 32 bita (ranije i 8).

MEMORIJA

Služi za skladištenje podataka u binarnoj formi. Svaki BAJT u memoriji ima svoju ADRESU. Memoriju (kao pasivni sklop) aktiviraju upravljački signali iz PROCESORA. Memorija se često deli na INTERNU i EKSTERNU.

Interna memorija

Obično je fizički realizovana u elektronskim kolima računara, "čipovima".

a) **ROM** (**R**ead **O**nly **M**emory)

je vrsta interne memorije čiji sadržaj obični korisnici mogu samo da čitaju. Najčešće se za ROM koriste specijalni memorijski čipovi - EPROM - u koje se sadržaj može upisivati samo uz pomoć specijalnih uređaja koji se nazivaju programatori EPROM-a. Kada isključimo računar sadržaj ROM memorije se **ne** gubi. U PC računarima ROM memorija se koristi, između ostalog, za smeštanje tzv. BIOS-a (**B**asic **I**nput **O**utput **S**ystem).

b) **RAM** (**R**andom **A**ccess **M**emory)

je vrsta interne memorije koju obični korisnik može ne samo da čita već i da menja njen sadržaj. Isključivanje računara dovodi do **brisanja** sadržine RAM-a. Radi se o memoriji koju programi koriste za svoj rad, zapravo na ovu memoriju se misli kada se govori o memoriji koja stoji na raspolaganju korisniku.

EKSTERNA MEMORIJA

je realizovana na nekom spoljašnjem nosiocu podataka (traka, disk...).

DISK

je magnetni medij spoljne memorije sa mogućnošću trajnog pohranjivanja informacija. Mi ćemo u ovom tekstu pod imenom disk podrazumevati harddisk, diskete i kompakt diskove.

a) **Harddisk**

je trvrdi nesavitljivi disk ogromnog kapaciteta. Najčešće je fiksiran i ne vadi se iz ležišta sem u slučaju zamene novim.

b) **Disketa** (Floppy disc)

je savitljivi disk koji se stavlja u disk jedinicu i vadi po potrebi. Kada je reč o PC računarima, diskete se mogu podeliti u dva tipa: 1) disketa prečnika 5.25 inča (sve se ređe koriste) sa kapacitetom od 360 KB ili 1.2 MB; 2) disketa prečnika 3.5 inča i kapaciteta 720 KB (sve ređe se koriste) ili 1.44 MB. Nešto ređe se koriste i **zipkete** koje su slične disketama od 3.5 inča ali su mnogo većeg kapaciteta.

c) **CD** (Compact Disc)

d) **DVD** (Digital Versatile Disc) je digitalni prilagodljivi disk ili digitalni video disk za smeštanje velike količine digitalizovanih podataka, npr. filmova.

PROCESOR

je elektronski sklop sposoban za izvršavanje pojedinih elementarnih operacija (čitanje, upis, sabiranje 2 broja..). Ukoliko su elektronska kola realizovana na jednom komadu poluprovodnika reč je o **MIKROPROCESORU**. Pomenute operacije sačinjavaju tzv. procesorski skup instrukcija. Taj set instrukcija je univerzalan i pomoću njih je moguće obaviti najraznovrsnije moguće operacije na binarnim podacima.

ASCII (American Standard Code for Information Interchange) kôd

ASCII /čita se ASKI) kôd je jedan od standarda kojim se (određenom permutacijom nula i jedinica) definišu skupovi znakova i vrši njihova numeracija od 0 do 255 (što je i najveći broj znakova koji se može specifikovati sa jednim bajtom, tj. osam bita memorije). Njime su obuhvaćena slova, brojevi, znakovi interpukcije, kontrolni kodovi, karakteristični znakovi evropskih jezika, grčka slova, grafički i specijalni znakovi).

II. OSNOVNI DELOVI PERSONALNOG RAČUNARSKOG SISTEMA (PC-a)

PC sistem se sastoji iz sledećih osnovnih delova:

1. CENTRALNE JEDINICE
2. MONITORA
3. TASTATURE I MIŠA
4. ŠTAMPAČA (NE NUŽNO)

1. CENTRALNA JEDINICA (Central Processing Unit - CPU) se sastoji od sledećih elemenata: a) osnovne ploče, b) spoljne memorije i c) napajanja. Smeštena je u kućište (Case) računara.

a) **Osnovna ploča (motherboard)** se sastoji od elektronskih komponenti neophodnih za rad PC-a. Na osnovnoj ploči se nalaze mikroprocesor, centralna memorija i magistrala (obavezno) i raznovrsne kartice za proširenje (opciono).

Mikroprocesori se mogu klasifikovati prema raznim kriterijumima, a najčešće je to s obzirom na:

- brzinu rada izraženu u tzv. radnom taktu (od nekoliko MHz do preko100 MHz),
- dužinu reči (8, 16 i 32 bita)
- proizvođača (INTEL i MOTOROLA su najpoznatiji, dok firme AMD, CYRIX, TEXAS INSTRUMENTS i sl. proizvode tzv. klonove INTEL-a). Procesori od pojave 80486DX i Pentium generacije mikroprocesora imaju u sebi ugrađen i matematički koprocesor.

Centralna memorija obuhvata čipove u kojima su realizovane ranije pomenute RAM i ROM memorije. Pored ove klasifikacije memorijskih čipova postoji i podela s obzirom na

- njihov kapacitet,
- brzinu rada,
- podela na statičke (brzi, ali malog kapaciteta i skupi - najčešće korišćeni za tzv. keš memorije) i dinamičke čipove (spori, zahtevaju tzv. "refreshment" sadržaja čipova, ali jeftini i velikog kapaciteta) i sl.

Magistrala (sabirnica, bus) je deo osnovne ploče koji služi za međusobnu komunikaciju svih komponenti PC-a. Magistralom prolaze bitovi koji nose najraznovrsniji sadržaj (od mikroprocesora ka periferiji i obratno), raznovrsni kontrolni signali, preko nje se ostvaruje sistem prekida i direktnog pristupa memoriji i sl. Zahvaljujući najnovijem tipu sabirnice, UBS (Universal Serial Bus) na računar je moguće preko USB porta (priključka) priključiti različite uređaje: miša, tastaturu, modeme, skenere, digitalne kamere, zvučnike... Na taj način prestaje potreba za korišćenjem tradicionalnih serijskih i paralelnih portova za priključivanje periferisjkih uređaja na računar.

Kartice za proširenje su elektronski sklopovi na kojima se bazira tzv. hardverska fleksibilnost PC-a. Na osnovnoj ploči su ranije postojali, kao standardni delovi, samo mikroprocesor, memorija i sabirnica. Sve ostalo što je potrebno korisniku priključivalo se na tzv. portove (slotove) za proširenje. Sada su na samoj osnovnoj ploći ugrađene i kartice različite namene.

Ove kartice se mogu podeliti na:
- memorijske kartice (omogućuju proširenje osnovnog RAM-a, čiji je kapacitet 1 MB - od čega je za operativni sistem i aplikativne programe rezervisano 640K - dodatnim količinama memorije)
- kontrolere za hard diskove i disketne jedinice (uređaji za upravljanje hardverom diskova i disketnih jedinica; u novije vreme su najpopularnija dva tipa kontrolera: IDE -Interface Design Enhancement- i SCSI -Small Computer System Interface-).
- video kartice (omogućuju realizaciju slike na ekranu monitora, određuju karakteristike i kvalitet te slike; postoje sledeći tipovi video kartica: CGA, MDA, HERCULES, EGA, VGA, SUPER VGA)
- kartice za komunikacije (npr. I/O - input/output za priključivanje štampača, za komunikaciju sa drugim računarima, kartice za povezivanje više PC-a u mrežu i sl.)
- ostale kartice (fax-modem kartica, kartica za skener, muzičke kartice, multimedia i sl.)

b) Standardnu **spoljnu memoriju** PC računara sačinjavaju tvrdi disk (jedan ili više) i disketne jedinice (jedna ili više). U novije vreme sve više su u upotrebi i sledeća dva medija masovne memorije: CD ROM i strimer (streamer) traka koji služe za pohranjivanje i prenos velikih količina podataka. Harddisk (tvrdi disk) je jedan od ključnih elemenata čitavog sistema: sa njega se učitava operativni sistem (najčešće), na njega se instalira aplikacioni softver, na njemu se skladište i sa njega se učitavaju podaci... Od njegovih performansi, dakle, zavisi kvalitet i brzina rada čitavog računara. Sledeće karakteristike ovog uređaja utiču na ukupan kvalitet rada na računaru:
- kapacitet
- tip disk interfejsa (tip komunikacije diska sa ostatkom PC-a)
- srednje vreme pristupa (broj milisekundi potreban za pronalaženje jednog sektora na osnovu zadatih koordinata)
- srednja brzina prenosa podataka (izražava se brojem prenesenih megabajta podataka u sekundi)
- prisustvo/odsustvo tzv. keša - softverskog ili hardverskog - (Softverski keš - deo RAM memorije u koji se upisuju jednom učitani podaci sa diska i koji mikroprocesor pretražuje pre sledećeg obraćanja disku - ukoliko tom prilikom naiđe na potreban podatak u kešu, preuzima ga odatle umesto da ga traži na disku - što drastično povećava brzinu pristupa podacima. Hardverski keš - logika rada je slična kao i u prethodnom slučaju, s tom razlikom što je ovaj keš fizički realizovan na samom disku/kontroleru).

2. **MONITOR** je osnovna izlazna jedinica računara.

3. **TASTATURA** i **Miš** su osnovne ulazne jedinice računara.

Generalno, što se hardverskih komponenti novijih PC računara tiče na delu je tzv. Plug& Play ("Utakni i Radi") tehnologija. Ova tehnologija koju podržava najveći broj novijih hardverskih komponenti od korisnika zahteva samo da umetne novonabavljenu komponentu na odgovarajuće mesto, tj. ležište, a **operativni sistem** računara ih po uključenju računara prepoznaje i uz pomoć odgovarajućih programa za upravljanje pojedinim uređajima (tzv. **drajvera**, engl. driver) omogućuje njihovo uključivanje u rad sistema.

III. OSNOVE RADA NA PERSONALNOM RAČUNARU POD OPERATIVNIM SISTEMOM WINDOWS

A. OPERATIVNI SISTEM

Operativni sistem je skup programa koji upravljaju resursima računara, rade sa diskovima, **folder**ima i **fajl**ovima i sl. Operativni sistem je posrednik između korisnika i resursa računara a stara se da korisničke zahteve koji su od korisnika direktno upućeni njemu ispuni angažovanjem raspoloživih hardvesrkih i softverskih resursa računara. Dakle, skup programa koji čine operativni sistem upravlja hardverom preko tzv. upravljačkih programa (**drajvera**), tj.komunikacionih mostova između operativnog sistema i pojedinih uređaja, a istovremeno brine o nesmetanom (i koordinisanom) odvijanju svih ostalih aplikativnih programa. Postoje različiti operativni sistemi (**DOS**, **OS/2**, **UNIX**, **XENIX**, **WINDOWS** itd.). Očigledno, budući da je operativni sistem posrednik kojem se korisnik neposredno obraća pri radu na računaru, svaki korisnik određenog računara mora poznavati način i pravila komunikacije sa operativnim sistemom pod kojim radi računar. Mada je doskora **DOS** (**D**isk **O**perating **S**ystem) bio preovlađujući operativni sistem PC računara trenutno je **WINDOWS** operativni sistem dominantan. Stoga će se u nastavku ovog teksta podrazumevati da računar radi pod operativnim sistemom **WINDOWS95** ili nekoj od novijih verzija ovog operativnog sistema (**WINDOWS98**). Za korisnike koji rade na računaru sa operativnim sistemom **DOS** ili **WINDOWS3.xx** , ako takvih još ima, postoji starija verzija ovog teksta koja se može dobiti od autora ovog dodatka. Premda ćemo u nastavku teksta objasniti i neke osnovne operacije koje će koristiti onima čiji računar radi pod verzijom **WINDOWS3.xx** (**xx** označava bilo koju podverziju ove verzije Windows-a) rad u Windows3.xx podrazumeva poznavanje **DOS**-a[1].

B. UKLJUČIVANJE I ISKLJUČIVANJE RAČUNARA POD OPERATIVNIM SISTEMOM WINDOWS

1. Uključivanje računara

Kada se sklopka (mrežni prekidač) POWER iz položaja OFF (ili 0) prebaci u položaj ON (ili 1) brujanje koje se čuje znak je da računar učitava u operativnu memoriju sve potrebne programe (ustvari OPERATIVNI SISTEM u širem smislu).

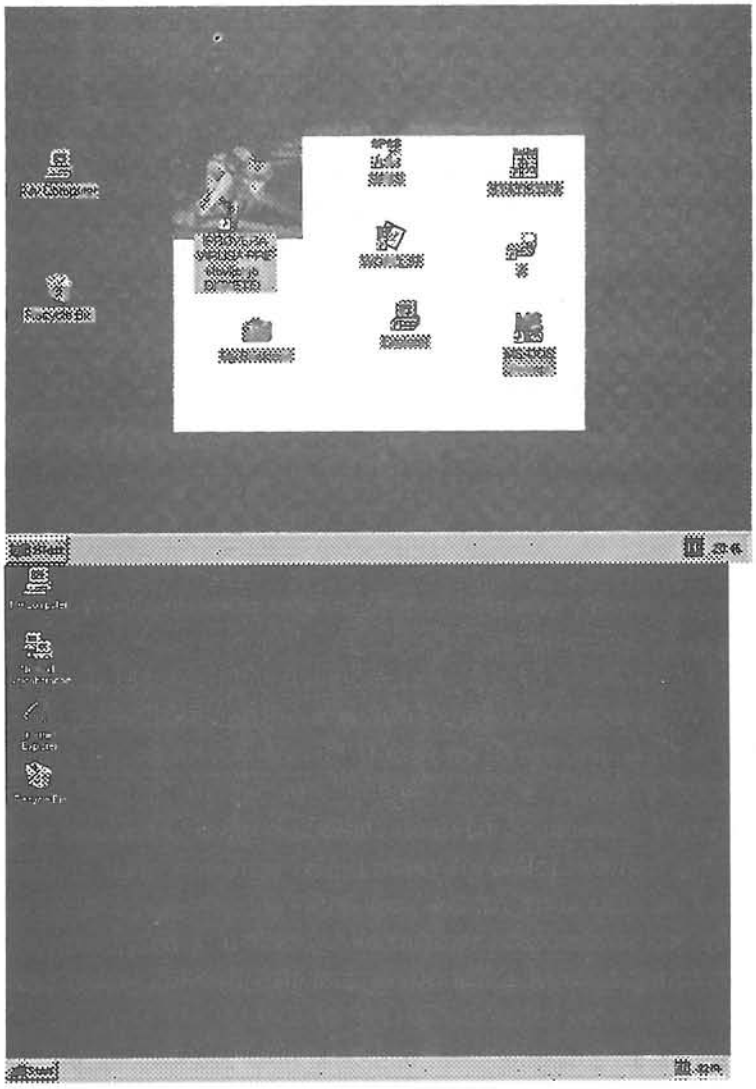
Prvo što se vidi na monitoru je poruka o verziji **BIOS**-a čime kompjuter informiše korisnika da učitava programe koji će njegove zahteve pretvarati u određene hardverske posledice. Sistemski **BIOS** (**B**asic **I**nput **O**utput **S**ystem) "podiže" računarski sistem, detektuje harddisk i upravlja osnovnim operacijama sa hardverom. BIOS na sistemima sa Plug&Play tehnologijom je posebno podešen tako da može da detektuje hardver i da upravlja promenama u konfiguraciji hardvera. Brojevi koji se menjaju vratolomnom brzinom govore o tome da računar proverava svoju memoriju. Kada se broj umiri on pokazuje koliko ukupno memorije stoji korisniku na raspolaganju za učitavanje aplikacionih programa i rad u njima. Po završenom poslu koji obavlja BIOS, učitava se operativni sistem koji preuzima dalji rad, kreira završnu konfiguraciju sistema, proverava da li postoje svi potrebni drajveri za postojeće uređaje i raspoređuje resurse pojedinim uređajem.

---

[1] Umesto xx u nazivu Windows 3.xx paketa različiti računari imaju različite brojeve, zavisno od verzije paketa.

Ukoliko računar radi u mreži, ili je na individualnom računaru to prethodnim konfigurisanjem predviđeno pre nego što se pojavi karakteristična **radna površina** (**desktop**) Windows-a može se pojaviti okvir za dijalog sa naslovom **Enter Windows Network** kojime se zahteva od korisnika da se prijavi, tj. "loguje" na računar. Prijavljivanje se sastoji u upisivanju (rad sa **tastaturom** objašnjen je u odeljku C ovog dodatka!) unapred definisanog korisničkog imena u polje **User Name** i lozinke, tj. tajne šifre koju "znaju samo korisnik i računar" u polje **Password**. Naravno, pri ukucavanju lozinke nije moguće videti njen sadržaj, zbog onih iza korisnikovih leđa koji možda baš tada napregnuto gledaju u ekran. Ako računar prepozna da korisničko ime i šifra odgovaraju nekome kome treba dozvoliti dalji rad, posle **klika** levim dugmetom miša (rad sa **mišem** objašnjen je u odeljku C ovog dodatka!) **OK** u datom okviru pojaviće se **radna površina**. Neprepoznati korisnik, ako nije sklon greškama u kucanju, možda pokušava da sedne za tuđ računar! U svakom slučaju, dokle god se proces prijavljivanja ne završi uspešno, nije moguće raditi na tom računaru.
Ako prijavljivanje pod korisničkim imenom nije predviđeno, radna površina se pojavljuje odmah po učitavanju operativnog sistema i završetka posla koji operativni sistem mora da obavi pri svakom uključivanju računara. To traje vrlo kratko!

Slika 1.Dva primera mogućeg izgleda radne površine pod Windows-om 95



Uočimo da ove dve radne površine imaju praktično istu liniju poslova (**taskbar**), tj. liniju na dnu ekrana. Na liniji poslova se pri uključivanju računara nalazi samo dugme **Start** (donji levi ugao), oznaka aktivnog jezika (plavi okvir u donjem desnom uglu na kojem su dva slova koja skraćeno označavaju jezik) i sat. Takodje, na radnoj površini oba prozora su ikonice **My computer** i **Recycle Bin**. Ostale ikonice prečica (**Shortcuts**), o čemu će biti reči kasnije, razlikuju se na dvema površinama jer to zavisi od toga koje programe korisnik najčešće koristi.

2. Isključivanje računara

a) W3.xx

Pre isključivanja računara potrebno je završiti rad sa Windows-om (bilo dvostrukim klikom na dugme kontrolnog menija u gornjem levom uglu Program menadžera, bilo odabirom **File** pa **Exit Windows** u meniju Program Manager-a). Pojavljuje se prozorčić sa naslovom Exit Windows i porukom **This will end Your Windows Session**. Potrebno je kliknuti na dugme **OK**. Kada se pojavi odzivni znak DOS-a (DOS prompt) isključiti računar.

b) W95

Odabrati **Start/Shut Down** opciju u meniju. Ukoliko nije automatski odabrana opcija. **Shut Down the Computer?** odabrati je (kliknuvši na belo dugme ispred ove opcije). Potom treba kliknuti na dugme **Yes** i sačekati da se pojavi poruka **It's now safe to turn off your computer**. **Tek posle pojavljivanja ove poruke na ekranu može se isključiti računar.** Ukoliko računar radi u mreži, pre opcije **Shut Down,** prema uputstvu na datoj mreži treba se odjaviti (opcija **Logg Off** ili **Disconnect**) sa mreže.

C. MIŠ I TASTATURA

1. Rukovanje mišem

Učenje rada sa mišem i prozorima najbolje je započeti prolaskom kroz Tutorial programa Windows. U Windows 3.xx (u daljem tekstu W3.xx) ovaj tutorijal se aktivira odabirom menija **Help** u liniji menija Program managera a potom stavke **Windows Tutorial**. U Windows95 (u daljem tekstu će za sve verzije Windowsa od Windowsa95 pa nadalje biti korišćena skraćenica W95) ovo možemo uraditi tako što odaberemo **Start/Help/** pa u okviru **Contents**-a Help-a **Tour: Ten minutes to using Windows** (nažalost za ovo Vam je, ukoliko ovaj program nije instaliran na harddisku potreban CD ROM).
Potrebno je naučiti sledeće radnje mišem:
1. Navigacija i nišanjenje (pointing) pokazivačem miša. Pomeranjem miša po podlozi pokazivač miša (strelica) se pomera po ekranu menjajući svoje koordinate. Miša ja najbolje držati tako da se poklopi celom šakom tako da kažiprst vrhom stoji na levom a domali (ili srednji, zavisno od debljine prstiju!) prst na desnom dugmetu.
2. Jednostruki klik (Click)-Brzo pritisnuti i odmah otpustiti levo ili desno dugme na mišu. Jednostruki klik na levo dugme miša najčešće služi selekciji, tj. markiranju određene stavke

(npr. fajla), a dvostruki klik levim dugmetom najčešće dovodi do otvaranja fajla ili foldera, ili, ako je reč o izvršnom fajlu, do aktiviranja programa koji se nalazi u tom fajlu.
3. Dvostruki klik (pritisak) (Double click)- Dva puta, u kratkom vremenskom razmaku, pritisnuti i otpustiti dugme miša.
4. Povlačenje (Drag) - nanišaniti željeni cilj, kliknuti na levo dugme miša i, **držeći pritisnuto levo dugme**, povući odabranu stavku na željenu poziciju. Kada smo stavku dovukli do željene pozicije otpustimo levo dugme.
5. Povlačenje desnim dugmetom (samo u W95) moguće je koristiti u nekim situacijama. U takvim slučajevima kada se povlačenje završi i otpusti desno dugme pojavljuje se meni u kojem se može odabrati šta treba uraditi sa prevučenom stavkom. Na željenu stavku u meniju se može kliknuti levim ili desnim dugmetom.

Zapamtite: U W3.xx praktično se jedino koristi **levo** dugme miša (ako je miš podešen za dešnjake), a u W95 levo i desno dugme. Desno dugme miša, zavisno od toga na šta je trenutno usmeren pokazivač miša, omogućuje veoma različite operacije. Obično je reč o skraćenim menijima kojima brže možemo obaviti neku operaciju koju je moguće zadati i na drugi, duži način. Da bi se otkrile sve funkcije koje pritisak na desno dugme miša ima u W95, najbolje je da korisnik sam pri različitim poslovima povremeno klikne na desno dugme miša i pogleda ponudjeni meni. Dugme u sredini, ako na mišu postoje tri dugmeta, se praktično nikad ne koristi.

2. Tastatura

(ovde je dat opis tzv. Enhanced tastature, 101)

Tasteri na tastaturi su organizovani u sledeće skupine:

a) Funkcijski tasteri F1 do F12 (gornji red, iznad alfanumeričkog dela - 12 tastera)

b) Alfanumerički deo (sredina - 58 tastera)

c) Numerički deo (desno - 17 tastera)

d) Prva grupa kontrolnih tastera (gornji red, desno od funkcijskih tastera - 3 tastera Print Screen, Scroll Lock, Pause)

e) Druga grupa kontrolnih tastera (6 tastera ispod prve grupe kontrolnih tastera)

f) Treća grupa kontrolnih tastera (4 tastera za kontrolu kurzora

g) Taster Esc koji stoji zasebno u gornjem levom uglu

a) Funkcijski tasteri imaju određene definisane funkcije kako u DOS-u tako i u drugim operativnim sistemima i programima (ali ne uvek ISTE!).
b) Alfanumerički deo ima slične tastere kao kucaća mašina ali i još neke dodatne: Ctrl, Alt, Enter,Win. Dugme Win, ako ga ima na tastaturi nalazi se između levog Ctrl i Alt dugmeta i ima na sebi stilizovanu slicicu prozora.

Zapamtite: Ako ste nekada kucali na pisaćoj mašini, uočite da **slovo O** i **cifra 0** NISU **ISTI,** KAO NI **slovo l** i **cifra 1!!!**

Shift — ukoliko se, pri kucanju teksta, drži ovaj taster i istovremeno pritisne neki drugi taster na ekranu se ispisuje, ako je u pitanju slovo - veliko slovo, ako su u pitanju brojevi, znaci interpukcije i drugi znaci - znak koji je ispisan u gornjem delu pritisnutog tastera.

Caps Lock — dokle god je uključeno, pri kucanju teksta, daje velika slova. Isključuje se i uključuje na isti način, pritiskom na istoimeno dugme. Kada je uključeno, signalnalampica Caps Lock u gornjem desnom uglu tastature zasvetli. Ne menja ispis tastera sa brojevima, znacima interpukcije i drugim znakovima.

Tab — pri kucanju teksta ima istu funkciju kao na pisaćoj mašini, uvlačenje kursora ili skokovito premeštanje kursora (tabulator). U radu sa prozorima za dijalog u Windows programima omogućuje prelazak sa jedne stavke na sledeću, iz jednog ulaznog polja u drugo i sl.

Ctrl — obezbeđuje svakom tasteru još po jednu funkciju. Najčešće korišćene kombinacije su: Ctrl X (za odsecanje selectovane stavke ili teksta-operacija CUT), Ctrl C (za kopiranje selektovane stavke ili dela teksta u bafer, operacija COPY) i Ctrl V (lepljenje odsečene ili kopirane stavke ili dela teksta na poziciju kurzora).

Alt — omogućuje, u kombinaciji sa odgovarajućom cifrom, unošenje ASCII kodova koji nisu na raspolaganju na tastaturi

Space — isto kao na pisaćoj mašini, razmaknica

Backspace — pri kucanju teksta služi za brisanje jednog znaka ulevo u odnosu na poziciju kursora.

ENTER — služi za završetak definisanja nekih komandi a pri kucanju teksta za namerni prelazak u sledeći red. Označava se još i sa -return- ili -cr- (od "carriage return" tj. povlačenja velike poluge na pisaćoj mašini čime se kolica vraćaju na početak novog reda).

c) Numerički deo:

Num Lock — definisanje uloge numeričkog dela tastature:
-ako je Num Lock uključeno (dakle, gori signalna lampica Num Lock u gornjem desnom uglu) numerička tastatura služi za kucanje brojeva, matematičkih operatora (za sabiranje +, za oduzimanje -, za množenje * i za deljenje /) i decimalne tačke.
-ako je Num Lock isključeno-numerički deo tastature može se koristiti za pomeranje kursora i kontrolne tastere. Budući da na tastaturama sada uglavnom postoji izdvojen kurzorski deo, najbolje je da dugme Num Lock bude stalno uključeno.

d) Prva grupa kontrolnih tastera

Prtsc	U Windows-u smešta sadržaj trenutnog ekrana u Clipboard (privremeni bafer). Ako se radi pod operativnim sistemom DOS, sadržaj trenutnog ekrana ide na štampač.
Scroll Lock	samostalan, nema nikakvu funkciju.
Pause	u nekim programima zaustavlja listanje sadržaja na ekranu.

e) Druga grupa kontrolnih tastera

Ins	prebacivanje iz Insert u Overwrite modus i obratno pri kucanju teksta.
Del	pri kucanju teksta služi za brisanje karaktera na kojem je kurzor ili markirane (selektovane) stavke.
Home	Postavljanje kurzora na početak tekuće linije pri kucanju teksta.
End	Postavljanje kurzora na kraj tekuće linije.
PgUp	U najvećem broju aplikacionih programa vodi na početak strane.
PgDn	U najvećem broju aplikacionih programa vodi na kraj strane.

f) Treća grupa kontrolnih tastera (kontrola kurzora).

Pritiskom na ove tastere kurzor se pomera u pravcu koji naznačuje strelica.

g) Esc	ima različite funkcije. U najvećem delu aplikacionih programa označava povratak na prethodni meni ili kraj rada.
h) Win	Ako ga ima na tastaturi, zamenjuje klik mišem na dugme Start na liniji poslova radne površine.

Zapamtite: Ovo su samo **najčešća** značenja dirki. U svakom programu treba naučiti njihova tačna značenja!

PREKIDI

Pritisak na tri tastera Ctrl Alt Del istovremeno (zapravo jedan pritisak na dugme DEL dok se dugmad Ctrl i Alt drže pritisnutima) omogućava "nasilno" prekidanje rada sa nekim programom. Po pritisku na ova tri dugmeta pojavljuje se prozor čiji je naslov **Close Program** i upozorenje Windows-a da će, ukoliko se ista kombinacija dugmadi pritisne još jednom računar biti resetovan (kao da uključujete ponovo računar ali bez testiranja memorije. To je zv. "meki (softverski) reset". Osim toga, u belom okviru ispod naslova dat je spisak programa koji trenutno rade.

Ukoliko neki od tih programa zaista želimo da prekinemo , treba klikom na levo dugme miša markirati, tj. selektovati naziv tog programa i zatim kliknuti na dugme **End Task**.Pri tome, treba znati da će sve što nije sačuvano na disku pre ovog prekida biti izgubljeno! Klik na dugme **Shut Down** omogućuje zaustavljanje svih aktivnih programa i isključivanje računara dok jednostruki klik levim dugmetom miša na dugme **Cancel** poništava dejstvo Ctrl Alt Del pritiska.
Pritiskom na dugme RESET na kućištu računara vrši se ponovno startovanje računara (potpuno isto kao kod ponovnog uključivanja). To je tzv. "tvrdi (hardverski) reset".
Ova dva prekida imaju najviše prioritete i njihovim aktiviranjem procesoru se nalaže da momentalno prekine ono što radi i pređe na instrukcije koje se nalaze u programu za opsluživanje datog prekida (tzv. interrupt handler).

Zapamtite: Prekidanje programa na ovaj način, kao i meki i tvrdi reset treba koristiti, **i to** navedenim redosledom, **samo kada se program "zaglavi" i bilo kakav pokušaj "odglavljivanja" programa ne pomaže.**

D. FORMATIRANJE DISKA

Da bi operativni sistem mogao da radi sa diskom potrebno je da ga najpre pripremi. Ova priprema se naziva **formatiranje** diska. Običan korisnik je ponekada u prilici da formatira diskete, ali nije gotovo nikad u prilici da formatira hard disk. Diskete se najčešće prodaju formatirane a i ukoliko se desi da korisnik pokuša da upotrebi neformatiranu disketu W95 će to prepoznati i predložiti formatiranje.
Formatiranjem se, između ostalog, na disk na određenim mestima upisuju oznake koje ga dele na površine koje se zovu SEKTORI. i STAZE (površine na disku koje se mogu zamisliti kao koncentrični krugovi). Prilikom čitanja ili upisivanja podataka na disketu glave disk jedinica se precizno nadnose nad staze. Nulta staza je na periferiji diska, dok je poslednja po broju najbliža centru. Kako svaka staza obuhvata određeni broj sektora, površina date diskete se može zamisliti kao izdeljena na n (n = broj staza x broj sektora) površina koje se nazivaju BLOKOVIMA. Toliko blokova ima i sa druge strane diskete (njima pristupa druga glava disketne jedinice). U svaki blok se može upisati određeni broj bajtova tako da je ukupni memorijski kapacitet diskete jednak proizvodu broja blokova i broja bajtova koji se mogu upisati u jedan blok. Međutim, najmanja jedinica memorije na disku koju operativni sistem može da evidentira nije blok, niti sektor, već KLASTER koji zavisno od diska obuhvata dva ili mnogo više blokova (na nekim tvrdim diskovima). Određen broj sektora tako formatiranog diska se koristi za sistemske potrebe, a ostatak za **fajl**ove. Sistemski deo svakog diska počinje od nultog sektora. Ovaj sistemski deo sadrži:
- Startni zapis
- FAT (**F**ile **A**llocation **T**able) tabelu (mapa prostora koji je dodeljen fajlovima; sadrži podatke o adresama sledećih klastera dodeljenih datom fajlu (adresa prvog klastera se nalazi u **dirketorijum**u ili **folder**u), kao i informacije o tome da li je klaster dodeljen ili slobodan, da li je ispravan ili nije i sl.
- Kopiju FAT tabele
- Osnovni (root ili korenski) folder, tj. direktorijum je tabela koja sadrži informacije o drugim direktorijumima i fajlovima. Sadrži ime, tip, atribute fajlova, vreme, datum njihovog kreiranja i poslednje izmene, adrese prvih klastera fajlova i njihovu veličinu.
Sve ove strukture sistemskog dela diska kreirane su upravo procedurom formatiranja.

Zapamtite: **Ukoliko se izvrši formatiranje diska na kojem postoje neki podaci, tada se celokupni prethodni sadržaj sa diska gubi!**

E. DRAJV

Drajv (engl. drive = nagon, pogon) je disk jedinica koja ima mehanizam za ČITANJE sa i UPIS informacija na disk.

Logičke oznake drajvova su :

A: - disketa(obično gornji diketni drajv ako postoje dva)
B: - disketa (obično donji disketni drajv)
C: - harddisk
D:, E: - oznake rezervisane za nove harddiskove, kompaktdisk ili nove particije istog tvrdog diska.

AKTIVNI (default) DRAJV-onaj drajv na kojem smo trenutno, tj. sa kojega računar trenutno čita ili na koji upisuje podatke.

F. FOLDER (DIREKTORIJUM)

Folder (engl. folder=fascikla) je spisak, tabela, katalog fajlova koji zauzimaju određeni logički prostor. Hijerarhijske je organizacije.
AKTIVNI folder je onaj u kojem smo! To je direktorijum na kojem operativni sistem podrazumeva da se nalaze fajlovi. Imena subfoldera se formiraju slično imenima fajlova (videti kasnije), s tom razlikom da se imenu subdirektorijuma ne pridružuje ekstenzija.

G. FAJL

Fajl (engl. file = datoteka) je skup informacija (program,tekst,podaci...) obuhvaćenih jednim IMENOM.

1. Ime fajla

Satoji se od OSNOVNOG IMENA (obavezno) i EKSTENZIJE (opciono). Osnovno ime i i ekstenzija se OBAVEZNO RAZDVAJAJU TAČKOM.

OSNOVNO IME FAJLA

Može sadržati od 1 do 255 karaktera proizvoljno odabranih. Ukoliko radimo pod operativnim sistemom DOS ili W3.xx onda osnovno ime sme imati najviše osam karaktera. Pri imenovanju fajlova nije moguće u istom folderu imati fajlove sa istim imenom. Sledeći karakteri nisu dozvoljeni kao deo imena fajla u W95: **: ' " / \ * ? |** . Ako koristimo DOS ili W3.xx onda nisu dozvoljeni ni sledeći znaci : **, [ ] > < + = ;**

EKSTENZIJA

Uobičajeno imati do 3 karaktera a u DOS i W3.xx najviše 3 karaktera. Govori o sadržini, tj. specifikuje tip datog fajla.
Primeri nekih standardnih ekstenzija :

.exe	izvršni program
.com	izvršni program
.dat	podaci u ASCII formatu
.txt	tekst
.c	izvorni program u C-u
.sys	konfiguracioni fajl, tzv. device driver
.ovl	deo veoma dugog programa
.ovr	deo veoma dugog programa
.bak	backup fajl
.bat	beč (batch) fajl
.html	HTML fajl (videti kasnije u delu teksta o Internetu)
.doc	tekst
.sav	fajl sa podacima u SPSS za Windows formatu
.spo	fajl sa ispisima u SPSS za Windows formatu
.sps	sintaksni fajl u SPSS za Windows programu

Zapamtite: Informacija o **imenu, dužini, datumu nastajanja i prostoru na disku** gde se fajl nalazi i čuvaju se u **direktorijumu.**

Zapamtite: Pri otvaranju i zapisivanju fajlova na disku treba voditi računa o specifikaciji fajla. **Specifikacija fajla** obuhvata:

- **oznaku disk jedinice** na kojoj je fajl (a:, b:, c:, d:, e:)
- **putanju foldera** ili direktorijuma u kojem je fajl, i
- **ime fajla**

Ukoliko se ne specifikuje disk jedinica podrazumeva se da je to aktivna disk jedinica, a ako se ne specifikuje folder podrazumeva se da je to aktivni folder.

2. Rezervisana imena

Pojedini uređaji ili jedinice računarskog sistema imaju u DOS-u rezervisana imena:

CON	-konzola (tastatura i monitor)
LPT1	-prvi paralelni port
LPT2	-drugi paralelni port
LPT3	-treći paralelni port
PRN	-isto kao LPT1
COM1	-prvi serijski port
COM2	-drugi serijski port
COM3	-treći serijski port
COM4	-četvrti serijski port
NUL	-nepostojeći uređaj
AUX	-isto kao COM1

Ova imena se zato ne mogu upotrebiti kao imena fajlova!

3. Izvršni fajlovi

Fajlovi sa ekstenzijama .COM .EXE .BAT (i još nekim) su izvršni fajlovi tj. njihov sdržaj čine programi koji se mogu izvesti.

4. Džoker znaci (wild cards)

Džokeri su karakteri koji može da zameni neki drugi karakter ili grupu karaktera u imenima fajlova ili foldera pri radu sa njima (ali ne i pri njihovom prvom imenovanju). Kao džoker znaci služe karakteri **?** i ***.**

Džoker ? zamenjuje karakter umesto kojeg stoji. Tako npr. ukoliko se zada naziv fajla ?ova.doc to se odnosi na sledeće fajlove: jova.doc, sova.doc, lova.doc.

Džoker * zamenjuje sve karaktere od onog umesto kojeg stoji pa udesno sve do tačke koja odvaja ekstenziju od imena (ukoliko fajl ima ekstenziju). Tako npr. ukoliko se zada ime fajla r*.txt to se odnosi na sledeće fajlove: roda.txt, r1994.txt, rotor.txt i sl.

H. OSNOVNE OPERACIJE U WINDOWS 3.xx I WINDOWS95

Premda će u nastavku ovog teksta glavna pažnja biti posvećena radu pod W95, objasnićemo i izvođenje pojedinih poslova u Windows-u 3.xx. Windows 3.xx nije pravi operativni sistem već pre programski paket koji je nadogradnja DOS-a i koji omogućuje izvršavanje aplikativnih programa. Da bi se iz DOS-a prešlo na rad u okruženju Windows-a 3.11 potrebno je na odzivni znak DOS-a (C:\>_) otkucati komandu WIN i pritisnuti dugme ENTER.[2]

Zapamtite: Ukoliko želimo da se Windows3.xx okruženje pojavi po uključivanju računara potrebno je kao poslednju komandu u AUTOEXEC.BAT fajlu upisati naredbu WIN. To, naravno, ne važi za WINDOWS95 jer se pri uključivanju računara automatski pojavljuje desktop (radna površina).

1. Rad sa prozorima (windows)

Sve operacije i programi se u windows okruženju obavljaju na radnom stolu (desktop). Pri startovannju Windows3.xx na radnoj površini se nalazi glavni prozor, glavna aplikacija unutar koje se izvode sve ostale aplikacije i poslovi, a to je Menadžer programa (**Program manager**). Iz njega se pokreću svi ostali programi.
U levom donjem uglu radne površine W95, po uključivanju računara, treba uočiti dugme START koje omogućuje odabir željene aplikacije ili posla. Na radnoj površini se nalazi nekoliko ikonica (My Computer, Recycle Bin /Korpa za otpatke/...) sa veoma važnim funkcijama. Na desnom kraju linije poslova nalazi se sat. Ako nanišanimo pokazivačem miša na broj koji pokazuje tačno vreme ubrzo će se pojaviti i informacija o datumu. Sa leve strane sata je, kako smo već pomenuli, plavo polje u kojem je oznaka trenutno aktivnog jezika. Klikom na to polje može se dobiti meni jezika koji su na raspolaganju tako da se pri pisanju tekstova u svakom trenutku promenom jezika može deo teksta pisati latinicom, deo ćirilicom a deo kineskim pismom. Naravno, pod uslovom da su jezičke podrške za slovenački, srpski i kineski jezik instalisane.

a) Programske grupe i programske stavke

Programske grupe su paketi programa za obavljanje odredjenih (srodnih) poslova, a programi unutar svake grupe se zovu programske stavke.
U W3.xx otvaranje programske grupe i programske stavke izvodi se tako što se pokazivačem miša nanišani ikonica programske grupe ili programske stavke (unutar programske grupe) i **dvaput** klikne na levo dugme miša
U W95 po kliku levim dugmetom miša na dugme Start na radnoj površini otvara se spisak programskih grupa. Kada se pokazivač miša pomeri na određenu programsku grupu i željeni program unutar te grupe, aktiviranje programa izvodi se **jednostrukim** klikom na levo dugme.

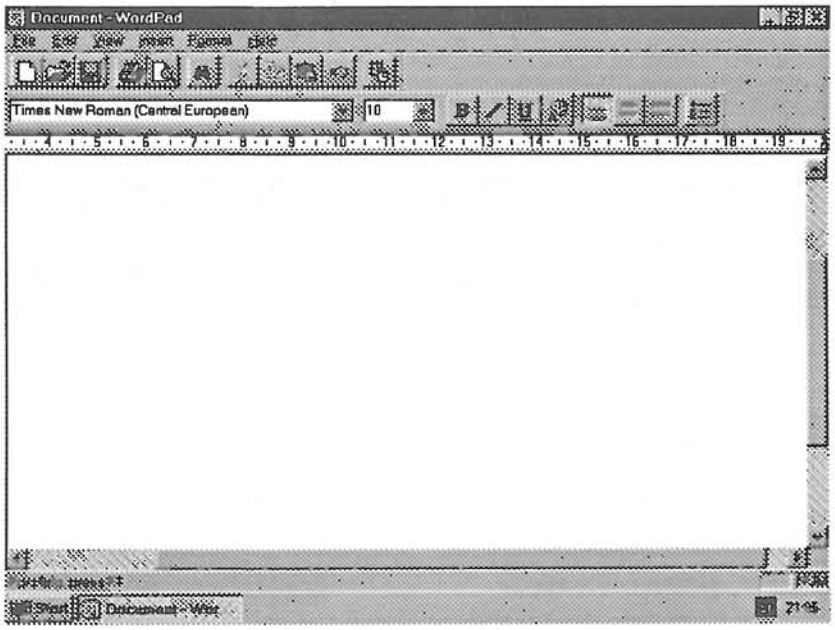
---

[2] Umesto xx u nazivu Windows 3.xx paketa različiti računari imaju različite brojeve, zavisno od verzije paketa.

b) Komponente prozora

Okvir prozora (Window border) služi za postavljanje strelice miša pri ručnoj promeni veličine prozora.
Naslovna linija (Title bar) sadrži naziv prozora. Naslovna linija aktivnog prozora je zatamnjena ili obojena (uobičajeno plavom bojom). Sa leve strane naslovne linije u Windows 3.11 je dugme kontrolnog menija (control menu box) a u desnom uglu dugmad za minimizovanje (strelica nadole-minimize buton) i maksimizovanje (strelica nagore-maximize buton) prozora. U Windows95, dugme kontrolnog menija predstavljeno je sličicom ili stilizovanim slovom (npr. na slici 2 je stilizovana ikonica programa za pisanje teksta WordPad), dugme za minimizovanje ima liniju na donjoj ivici a dugme za maksimizovanje na gornjoj ivici. Na slici dva ne vidi se dugme za maksimizovanje jer je prozor već prethodno maksimizovan. Treće dugme u desnom gornjem uglu naslovne linije (precrtano putačom) postoji samo u W95 i služi za zatvaranje prozora.Treba razlikovati **minimizovanje** i **zatvaranje** prozora. Pri minimizovanju prozora ono što je u prozoru (program, fajl...) je još uvek aktivno a pri zatvaranju prozora programi se okončavaju a fajlovi zatvaraju. Pri minimizovanju prozora aktivnog programa u W3.xx se pojavljuje sličica aplikacije (application icon) koja je nešto drugačija od ikone zatvorenog prozora. U Windows95 se nazivi minimizovanih prozora prozora nalaze na liniji poslova (task bar) koja se nalazi na dnu radne površine. Na slici 2 se na liniji poslova nalazi naziv *Document-Wor..* jer je aktiviran trenutno jedino program WordPad.

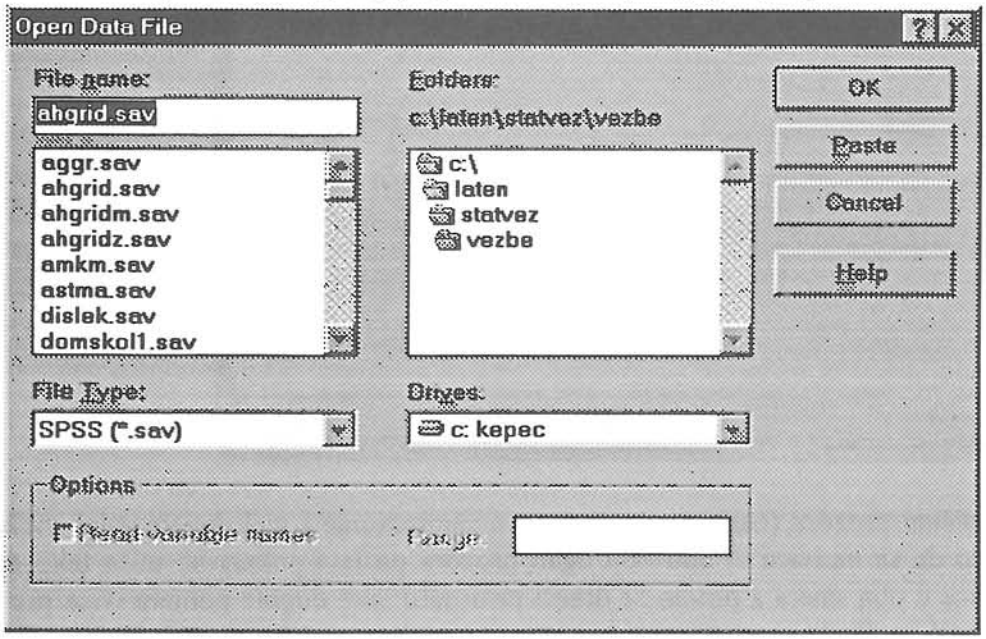
Slika 2: Izgled jednostavnog prozora u W95



Promena veličine prozora (osim dugmetom za minimizovanje i maksimizovanje) može se uraditi i tako da se na ivicu ili bilo koji ugao prozora postavi pokazivač miša tako da se pojave strelice u oba smera a potom se držeći pritisnuto levo dugme pomera ivica prozora do željene veličine. Kada se prozor poveća, u gornjem desnom uglu se pojavljuje, u W3.xx dugme sa dvostrukom strelicom (gore i dole), a u W95 dugme sa dvostrukim kvadratićem (kao na slici 2) za obnavljanje tj. vraćanje prozora na prethodnu veličinu (restore buton).

Linija sa menijima (Menu bar) prozora nalazi se neposredno ispod naslovne linije i sastoji se od odredjenih reči koje čine stavke menija (File, Edit, View, Window, Help i sl.). Klikom levim dugmetom miša na datu reč u meniju otvaraju se (poput roletne) opcije date stavke.
Ikonice alatki ili Sličice (**Icon bar** ili **Tools bar**) nalaze se u nekim prozorima pojedinih programa. Pritiskom na odredjenu sličicu odabira se odredjena komandna stavka datog programa. Npr. Umesto odabira stavke File pa potom opcije Print moguće je zatražiti štampanje fajla u aktivnom prozoru klikom na ikonu, tj. sličicu štampača unutar datog prozora.
Klizači (Scroll bars) su trake sa desne (vertikalni klizač) i donje (horizontalni klizač) strane prozora koje su ograničene strelicama sa obe strane a sadrže pravougaono ili kvadratno dugme na traci. Klizači služe za listanje informacija koje ne mogu da stanu istovremeno u dati prozor. Pomeranje sadržaja prozora vrši se bilo klikovima na strelice, bilo klikom levog dugmeta na kvadratno dugme i potom (uz držanje stiska na levo dugme) povlačenjem kvadratića u željenom smeru.
Prozor za dijalog (Dialogue box) se pojavljuje u trenucima kada korisnik treba da unese dodatne informacije o zadatku koji treba obaviti. Na slikama 3 i 4 su dati primeri prozora za dijalog iz programa SPSS, verzija 6.1). U dijalog prozoru se nalaze **ulazna polja** (text boxes -na slici 3 to je polje ispod File name: u kojem piše ahgrid.sav) u koja se upisuje potrebna informacija, ponekad **prozorčići sa spiskovima** (list boxes- na slici 3 prozorčić sa leve strane sa spiskom fajlova: aggr.sav ...), **prekidači** (Check boxes) - klikom na prekidač (okruglo ili kvadratno dugme) ispred imena opcije odabira se ili poništava data mogućnost, **dugmad za izbor odgovarajućih opcija** (Option buttons-na slici 4, dugmad pored reči Means, Sum ...ili pored reči Ascending Means...) , liste sa fiksnim sadržajem (Drop-down list boxes-na slici 3 to je polje ispod natpisa Drives:) i komandna dugmad (Npr. **Ok** - za završetak dijaloga, **Cancel** - za odlazak iz dijaloga i odustajanje od datog zadatka i **Help**- za pozivanje pomoći).

Slika 3: Primer prozora za dijalog (koji služi otvaranju fajla) iz programa SPSS



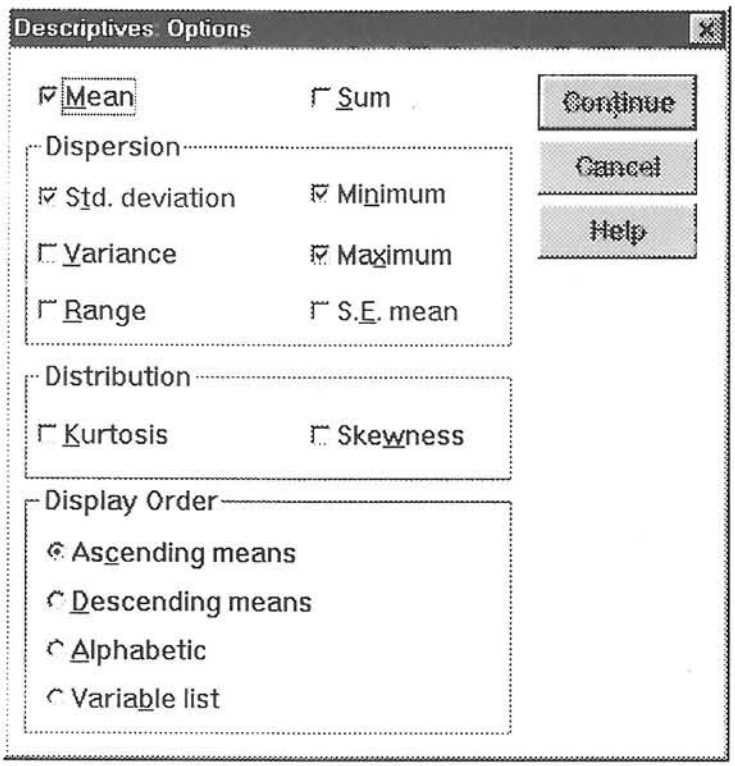
Zapamtite: Prelazak iz jednog polja u drugo unutar prozora za dijalog može se vršiti i pritiskom na dugme **Tab** (ili u suprotnom smeru **Shift Tab**) na tastaturi.

c) Zatvaranje prozora

U W3.xx zatvaranje prozora se vrši dvostrukim klikom na kontrolno dugme u gornjem levom uglu ili jednim klikom na kontrolno dugme i klikom na opciju **Close** u tako otvorenom meniju.

U W95 se zatvaranje prozora (osim dvostrukim klikom na sličicu u gornjem levom uglu prozora slično kao u W3.xx) može postići i jednostrukim klikom na dugme sa putačom u gornjem desnom uglu prozora.

Slika 4. Primer jednog dela prozora za dijalog iz programa SPSS



Zapamtite: Zatvaranje prozora nije isto što i njegovo **minimizovanje.** Zatvaranjem prozora u kojem radi neki program prekidamo rad tog programa, dok **u minimizovanom prozoru program i dalje radi !**

Zapamtite: Pri bilo kojem započetom definisanju operacije koja treba da se izvrši, klik na dugme **Cancel** dovodi do poništavanja započete definicije i vraćanja na prethodni nivo.

d) Rad sa više prozora istovremeno

I u W3.xx i u W95 moguće je istovremeno imati otvoreno više prozora. Svi otvoreni prozori mogu po želji biti poredjani i kaskadno (jedan iza drugog) ili kao pločice (jedan pored drugog). Kaskadni raspored otvorenih prozora u Windows3.11 postiže se odabirom opcije **Cascade** u meniju **Window** a pločasto odabirom opcije **Tile**. Jedan prozor je uvek trenutno aktivan a drugi neaktivni. Iz prozora u prozor unutar istog programa se može ići bilo klikom unutar prozora u koji se želi ići bilo odabirom naslova prozora u meniju Window.
U W95 naslovi svih otvorenih prozora (od kojih su neki prozori aktivnih programa) nalaze se na liniji poslova (Task bar) na donjoj ivici radne površine i prelazak iz jednog prozora u drugi može se vršiti i klikom levog dugmeta na naslov prozora u koji se želi ići. Prozor u koji "prelazimo" postaje **aktivni** prozor što se vidi po zatamnjenosti njegove naslovne linije. Promena načina raspoređivanja otvorenih prozora u W95 postiže se tako što se pokazivač, tj. strelica miša postavi na **prazan** prostor linije poslova (Task bar), klikne na desno dugme i (u tako otvorenom meniju) klikne na željenu opciju: Cascade, Tile Horizontally, Tile Vertically.

2. Aktiviranje programa u Windows okruženju

a) W3.xx

Aktiviranje programske grupe i programske stavke u W3.xx-vrši se dvostrukim klikom levog dugmeta miša na ikonicu programske grupe a potom (na isti način) odredjene programske stavke. Pri aktiviranju odredjene programske grupe ili stavke otvara se prozor pokrenute aplikacije.

b) W95

Ukoliko na radnoj površini postoji ikonica željenog programa on se aktivira dvostrukim klikom levog dugmeta miša na ikonicu. Ako nema ikone željenog programa na radnoj površini, treba kliknuti na dugme START u donjem levom uglu prozora, pomeriti pokazivač miša na programsku grupu, tj. na programsku stavku unutar programske grupe, a onda kliknuti (levim dugmetom) na reč koja predstavlja dati program.
Npr. START/SPSS/SPSS

Zapamtite: Ukoliko želimo da prelazimo iz jednog aktivnog programa u drugi, to je najlakše uraditi kombinacijom tastera **Alt Tab.** Treba **držati pritisnuto** dugme **Alt** na tastaturi a dugme **Tab** pritiskati dok se ne pojavi ime programa u koji hoćemo da odemo. U trenutku kada se pojavi **ime** željenog programa treba **otpustiti** dugme **Alt.**

3. Rad sa diskovima, folderima i fajlovima u Windows okruženju

Za kopiranje, brisanje, promenu imena i druge operacije sa diskovima, direktorijumima i fajlovima u W3.11 koristi se program **File Manager** u okviru programske grupe **MAIN**. Po mišljenju pisca ovih redova, u slučaju da ste prinuđeni da radite pod W3.xx, bolje je naučiti komande DOS-a kojima se ove operacije mogu obaviti nego prepustiti se neiscrpnim maltretiranjima od strane **File Manager**-a.
U W95 za rad sa diskovima, folderima i fajlovima može se koristiti program **Windows Explorer** koji se aktivira odabirom **Start/Programs/Windows Explorer.**

a) Prelazak sa diska na disk

Po aktiviranju Windows Explorera, u levoj polovini njegovog prozora, pojavljuju se ikonice sa logičkim oznakama drajvova koji postoje na računaru (A:, C: i sl.). Na željeni disk prelazimo dvostrukim klikom na logičku oznaku drajva u kojem je disk. U desnom okviru Eksplorerovog prozora će se tada pojaviti prikaz (hijerarhijske) strukture foldera na disku na koji smo prešli i imenafajlova koji se nalaze u osnovnom folderu diska. Foldere i fajlove je veoma lako razlikovati jer ikonice foldera liče na fascikle sa širom horizontalnom no vertikalnom stranom. Ukoliko pored nekog foldera stoji sa desne strane znak + to znači da taj folder ima i subfoldere koje trenutno ne vidimo.

b) Prelazak iz foldera u folder

Dvostruki klik levim dugmetom na ime ciljnog foldera vodi nas u taj folder a ikonica foldera se "otvara". U desnom okviru Eksplorera pojavljuju se imena subfoldera i fajlova koji se nalaze u otvorenom folderu.

c) Selekcija ili markiranje foldera i fajlova

Odabir, tj. markiranje pojedinačnih foldera i fajlova vrši se jednostrukim levim klikom na ime foldera ili fajla. Poništavanje selekcije se postiže klikom mišem u prazan prostor.
Ukoliko su folderi ili fajlovi koje želimo da markiramo jedan ispoddrugog u neprekinutom nizu sve njih možemo markirati odjednom tako što kliknemo na ime prvog foldera ili fajla, premestimo pokazivač miša na poslednji i, **prethodno pritisnuvši** dugme **Shift** na tastaturi kliknemoposlednji folder ili fajl u spisku. **Dugme Shift za vreme poslednjeg klika moramo držati pritisnutim.**
Za markiranje više fajlova ili foldera, koji nisu u neprekinutom spisku, tj. na "preskok" treba pritisnuti dugme **Ctrl** na tastaturi i kliktati redom na željene fajlove. Za sve to vreme dugme Ctrl se mora držati pritisnutim.

d) Otvaranje fajlova u Windows Exploreru

Ukoliko fajl nije izvršni, tj. programski dvostrukim klikom na ime fajla možemo fajl otvoriti i videti njegov sadržaj. W95 sam poziva program u kojem je taj fajl **kreiran** (ako takvog programa ima na računaru). Ukoliko se neizvršni fajl ne može otvoriti na ovaj način onda moramo aktivirati program koji može da otvori fajl pa iz tog programa otvoriti fajl (videti objašnjenje pod tačkom **j** u ovom delu teksta). Ako je reč o izvršnom fajlu, dvostruki klik na ime tog fajla dovešće do aktiviranje programa sadržanog u fajlu.

Zapamtite: Imena foldera i fajlova u W95 može (za razliku od DOS-a i W3.xx) mogu biti duga do 255 slovnih znakova. Ipak, ako fajlove čije je osnovno ime duže od osam znakova budete koristili na računaru koji radi pod DOS-om ili W3.xx moći ćete da vidite samo njihova skraćena imena. Da biste, dok još radite pod W95, videli kakvo će ime imati takav fajl u DOS-u ili W3.xx treba u prozoru **Windows Explorera** u kojem je spisak fajlova nanišaniti na ime fajla, kliknuti desnim dugmetom i u meniju odabrati opciju **Properties** a u rezultatu potražiti rubriku **MS-DOS** pod jezičkom **General.**

e) Kopiranje fajlova i foldera unutar istog diska

VarijantaI: Po markiranju fajla ili foldera za kopiranje pritisnuti dugme **Ctrl** na tastaturi i, **ne otpuštajući dugme Ctrl**, povući odabranu stavku na ime foldera u koji želimo da je kopiramo. U trenutku kada se nanišan ciljni folder otpustiti levo dugme miša i dugme Ctrl.
VarijantaII: Kada je stavka koja se kopira markirana u meniju Explorera odabrati **Edit/Copy**, preći u ciljni folder a potom odabrati u meniju **Edit/Paste**.

f) Kopiranje fajlova i foldera sa jednog diska na drugi

VarijantaI: Po markiranju fajla ili foldera za kopiranje povući odabranu stavku na ime foldera, **unutar drugog diska**, u koji želimo da je kopiramo. U trenutku kada se nanišan ciljni folder otpustiti levo dugme miša.
VarijantaII: Kada je stavka koja se kopira markirana u meniju Explorera odabrati **Edit/Copy**, preći na ciljni disk i u ciljni folder a potom odabrati u meniju **Edit/Paste**.

Zapamtite: Ako želimo da KOPIRAMO fajlove iz foldera u folder NA ISTOM DISKU njihovim povlačenjem mišem iz izvornog u ciljni folder treba **držati za vreme povlačenja** dugme **Ctrl** na tastaturi. Ukoliko povlačenje odabranih falova vršimo bez držanja Ctrl dugmeta fajlovi će biti PREMEŠTENI (a ne kopirani!) iz izvornog u ciljni folder.
KOPIRANJE FAJLOVA SA JEDNOG DISKA NA DRUGI može se izvesti prostim povlačenjem mišem odabranih fajlova. Ako hoćemo,pak, da fajlove PREMESTIMO sa izvornog na ciljni disk (a ne da ih samo kopiramo) onda prilikom povlačenja treba **držati** taster **Shift.**

g) Pravljenje novog foldera

VarijantaI: Kliknuti na ikonicu **My Computer** radne površine W95 a potom na ikonicu harddiska ili diskete i direktorijuma (foldera) ispod kojeg se želi praviti novi folder. Odabrati stavku **File/New/Folder** a potom dati ime folderu i pritisnuti Enter.
VarijantaII: U Windows Exploreru, po markiranju foldera ispod kojeg želimo da napravimo novi subfolder treba u meniju **File** odabrati **New Folder** i po završenom upisivanju imena novog foldera pritisnuti dugme **Enter** na tastaturi. Ukoliko ne upišemo ime novog foldera W95 će folderu dati ime **NewFolder**.

h) Promena imena postojećeg foldera ili fajla

Varijanta I: Po markiranju foldera ili fajla kojem želimo da damo drugo ime, u meniju Explorera odaberemo **File/Rename**. U tom trenutku okvir oko postojećeg imena kao da se "opusti". **Nije potrebno brisati staro ime jer će ceo markirani tekst automatski nestati kada počnemo da ukucavamo novo ime!** Po završenom upisivanju novog imena fajla ili foldera pritisnuti dugme **Enter** na tastaturi ili kliknuti mišem u prazan prostor prozora Explorera.
Varijanta II: Po markiranju foldera ili fajla kojem želimo da damo ime, kliknemo **jedanput** levim dugmetom miša na postojeće **ime**. U tom trenutku okvir oko postojećeg imena kao da se "olabavi". Umesto starog umena treba upisati novo i po završenom upisivanju novog imena novog foldera pritisnuti dugme **Enter** na tastaturi. **Razmak između jednostrukog klika za markiranje fajla i ponovnog klika na ime fajla mora biti dovoljno dug da W95 ta dva klika ne razume kao dvostruki klik!**

i) Brisanje postojećeg foldera ili fajla

Po markiranju foldera ili fajla koji želimo da brišemo pritisnuti dugme **Del** na tastaturi ili ime fajla ili foldera povući do ikonice **Recycle Bin** i otpustiti. U tom trenutku, ako se fajl ili folder koji se brišu nalaze na harddisku, W95 postavlja pitanje **Are you sure you want to send ...to the Recycle Bin**? Ukoliko kliknemo na dugme **Yes** fajl će biti privremeno izbrisan sa harddiska, tj. biće premešten u korpu za otpatke (Recycle Bin). Iz korpe za otpatke fajl možemo oživeti sve dok postupak ne ponovimo u folderu **Recycle Bin**. Ako otvorimo korpu za otpatke (recimo dvostrukim klikom na ikonicu Recycle Bin na radnoj površini W95 ili u levom okviru Windows Explorera) i fajl izbrišemo iz Recycle Bin-a, fajl će biti izbrisan zauvek. Fajlovi i folderi koje brišemo sa diskete ne premeštaju se u Recycle Bin i biće u prvom krugu trajno izbrisani.

j) Otvaranje postojećeg fajla iz aktivnog programa

Kada smo aktivirali neki program, fajlove koje taj program može da čita otvaramo tako što u liniji menija datog programa odaberemo **File/Open** i zatim specifikujemo disk (ulazno polje **Drives:**), folder (dvostrukim klikom na folder u kojem je fajl koji otvaramo) i ime fajla

(dvostrukim klikom na ime fajla) koji želimo da program otvori. Specifikovanja diska vrši se tako da se klikne na strelicu na desnom kraju ulaznog polja **Drives** -pogledati sliku 3) i u padajućoj listi koja se pojavi odabere disk klikom na logičku oznaku odgovarajućeg drajva. Specifikovanje foldera izvodi se tako što se u strukturi foldera klikne dva puta na ime željenog foldera.

k) Kreiranje i čuvanje (zapisivanje na disku) novog fajla u nekom od programa

Kada smo aktivirali neki program, novi fajl kreiramo tako što u liniji menija datog programa odaberemo **File/New**. Kada u fajl upišemo ili unesemo željeni sadržaj možemo ga sačuvati odabirom **File/Save**. Potrebno je da zatim u prozoru za dijalog specifikujemo disk (ulazno polje **Drives:**) i folder (dvostrukim klikom na folder u kojem je fajl koji otvaramo) u koji ćemo fajl zapisati kao i ime fajla (upisivanjem imena fajla u polje **File name**). Operacija definisanja čuvanja fajla na disku završava se, zavisno od programa, klikom na dugme **Save** ili **OK**.

l) Čuvanje (zapisivanje na disku) otvorenog fajla u nekom od programa pod drugim imenom

Dok je fajl otvoren, u meniju programa treba odabrati **File/Save as** (SAČUVAJ KAO) i pošto se specifikuje disk (ulazno polje **Drives:**) i folder (dvostrukim klikom na folder u koji treba zapisati fajl) upisati novo ime fajla u polje **File name**. Operacija definisanja završava se, zavisno od programa, klikom na dugme **Save** ili **OK**. Ukoliko pri tome menjamo i tip fajla u polju **Save As Type** treba odabrati odgovarajući tip.

m) Štampanje sadržaja fajla iz aktivnog programa

Dok je fajl otvoren, u meniju programa treba odabrati **File/Print** (Štampaj). U dijaloškom prozoru **Print** treba definisati opcije ako one koje su ponuđene ne odgovaraju i kliknuti na dugme **OK.** Podrazumeva se, naravno, da je štampač već uključen i snabdeven papirom.

n) Prečice (Shortcuts)

Da bismo aktivirali neki program koji vrlo često koristimo ili pregledali neki folder koji stalno pregledamo, moguće je napraviti prečicu kojom zamenjujemo uobičajenu proceduru aktiviranja programa ili pregledanja nekog foldera. Ikonu prečice možemo smestiti na radnu površinu (na slici 1 mogu se videti, npr. ikone prečica za razgledanje foldera **Recycle Bin** i prečice za aktiviranje programa **Internet Explorer**).
Dvostruki klik na kreiranu prečicu zamenjuje uobičajenu i dužu proceduru aktiviranja programa ili odlaska u neki folder.
Postupak kreiranja prečica i njihovog smeštanja na radnu površinu može se savladati pomoću **Help**-a W95.

4. Aktiviranje programa koji rade pod DOS-om u Windows95 okruženju

Programe koji rade pod DOS-om moguće je aktivirati bar na tri načina:

a) Odabrati **Start/Run** i u ulaznom polju Open prozora Run upisati specifikaciju (putokaz i osnovno ime izvršnog fajla) fajla u kojem je program. Npr.za pozivanje Norton Editora treba upisati C:\UTL\NED ako je izvršni fajl NED.COM u folderu \UTL na harddisku.

b) Kliknuti na ikonicu **My Computer** (na desktopu) ili aktivirati program Windows Explorer. Potom "otići" na disk i u folder (kako je objašnjeno ranije) u kojem je izvršni fajl. Kliknuti dva puta levim dugmetom miša na ime izvršnog fajla.

c) Odabrati **Start/Programs/MS-DOS Prompt**. Po privremenom izlasku u DOS na odzivni znak DOS-a pozvati izvršni fajl (na način kako se to inače radi u operativnom sistemu DOS). Za napuštanje DOS-a (posle privremenog izlaska i završetka rada sa DOS-om) i zatvranja prozora u kojem je DOS radio treba kao komandu DOS-a otkucati **Exit** (i, naravno, pritisnuti dugme Enter).

> **Zapamtite:** Traženje odredjenog fajla ili grupe fajlova (videti o džoker znacima u odeljku III.G.4) se u W3.11 vrši korišćenjem File Manager-a a u W95 odabirom **Start/Find/Files or Folders**

IV. OSNOVE KORIŠĆENJA INTERNET EKSPLORERA

A. INTERNET

Internet je zapravo "mreža svih mreža", globalna svetska mreža nastala medjusobnim povezivanjem (internetworking = medjumrežno povezivanje) računarskih resursa širom sveta[3]. Njeni zameci su nastali u hladnoratovsko vreme kao deo projekta čiji cilj je bio da poveže računare američkog Ministarstva odbrane radi održavanja visokog nivoa komunikacije u slučaju nuklearnog udara (СПАСИБО Rusima!). Danas ovu mrežu čini na desetine miliona računara a svaki od njih ima svoj IP (Internet Protocol), broj na osnovu kojega mu svaki drugi računar može poslati informacije. S obzirom na ogromnu količinu informacija kojima se (ukoliko ste uključeni u ovu "mrežu svih mreža") može pristupiti, pravi problem zapravo nije da li neka informacija koja vam je potrebna negde postoji, već kako do nje doći u razumnom vremenskom roku.

---

[3] Tekst u odeljku **A** je deo teksta koji je autor dodatka objavio u Psihološkim novinama, br.25. Veliku zahvalnost za pomoć u pisanju ovog dela teksta dugujem kolegi **Daliboru Košutiću** čije je nesebično zalaganje doprinelo da studenti psihologije imaju mogućnost da praktično upotrebe znanja kojima treba da ih nauči ovaj Priručnik u celini.

Tri glavna servisa, tj. usluge koje postoje na Internet-u su:

a) Elektronska pošta (E-mail) ;

b) World Wide Web i

c) Usenet/News ili konferencije.

a)Elektronska pošta je servis Interneta, koji po svojim principima funkcionisanja potpuno odgovara klasičnoj pošti. Jedino je u odnosu na klasičnu (ili, kako se to sad popularno zove "snail mail" tj. "puževa pošta") neuporedivo brža i jeftinija. Osim toga, ako poštari štrajkuju vi i pored toga možete slati i primati (elektronska) pisma.

b)World Wide Web (u daljem tekstu **WWW**) su zapravo nepregledne skupine "oglasnih tabli" okačenih širom sveta koje samo čekaju da "stanete ispred njih" i pročitate sve što vas zanima. Ovde zaista važi ona vašarska "nema šta nema". Istina, za neke od tih informacionih servisa (npr. pregledanje pojedinih biblioteka) potrebno je ne samo vreme već i novac.

c)Usenet/News ili konferencijski sistem je Internet-ov servis koji opslužuje ljude iz celog sveta ali u vezi sa nekom specifičnom, odredjenom krugu korisnika zajedničkom temom (npr. socijalne navike pingvina ili najbolji način da se ukiseli kupus).

B. ZNAČENJE NAJVAŽNIJIH TERMINA U VEZI SA INTERNETOM

Provajder je organizacija ili entitet koji korisnicima obezbedjuje pristup Internetu ;

Link je beočug u mreži Interneta ili (unutar odredjene prezentacije na WWW) putokaz, tj. usmerivač sa date lokacije na neku drugu;

Server je računar koji sadrži odredjenu prezentaciju i koji ima odredjeno ime (npr. Beograd.com);

Nalog je dozvola, tj. registracija na osnovu koje korisnik može pristupati mreži;
URL (Uniform Resource Locator) predstavlja adresu na kojoj se u okviru WWW nalazi neki dokument, tj. prezentacija. URL adresa ima tri osnovna elementa:

- **Protokol** ili princip preuzimanja. Najčešći protokoli su **HTTP** (Hyper Text Transfer Protocol) i **FTP** (File Transfer Protocol).
- **Lokacija** na mreži definiše mrežni server ili Internet host na kojem je dokument smešten.
- **Putanja** do resursa ili konkretna lokacija prezentacije na zadatom serveru.

Finger - servis Interneta koji daje informacije o trenutno prijavljenim korisnicima na računaru;

Telnet- servis Interneta koji, uz uslov da se ima odgovarajući nalog, omogućuje povezivanje korisnika na neki udaljeni računar i dalji rad na tom računaru kao da je na radnom stolu korisnika;

Pretraživač - server koji vrši pretragu prema zadatom kriterijumu i locira tražene podatke. Najpoznatiji pretraživači na Internetu su **Yahoo!** i **AltaVista**.

HTML (Hyper Text Markup Language) - jezik u kome se pišu WWW prezentacije. Dokumenti napisani u ovom jeziku osim teksta i specijalnih sekvenci za formatiranje sadrže i ukazivače (obično podvučene reči ili delovi slika) na neke druge dokumente (koji se nalaze negde u mreži, često na udaljenim serverima) a u kojima se podrobnije opisuje pojam ili deo slike.

Hipermedijalni dokument - dokument u kome je tekst kombinovan sa slikama, grafikom, zvukom, animacijom i u kojem se iz bilo kojeg medijuma može pozvati neki drugi medijum (iz teksta zvuk, iz zvuka animacija i sl.).

Home page - Matična strana, tj. mesto na kojem se na Interentu nalazi odredjena prezentacija.

Web site- Lokacija ili mesto na Internetu.

C. INTERNET EXPLORER

Po aktiviranju programa, u ulazno polje **Address:** (slika 5) treba uneti odgovarajuću **URL** adresu koja dovodi do otvaranja određene WWW stranice ili HTML dokumenta sa sopstvenog računara ili mrežnog servera.

Slika 5: Prozor Internet Explorera

U slučaju da želimo da otvorimo fajl sa svog računara ili mrežnog servera u polje **Address:** unosimo **putanju** (tj. specifikaciju) fajla u sledećem obliku:

logička_ oznaka_ drajva\folder\ime fajla (za sopstveni računar)
(Npr. C:\Slike\grafika.html)
ili

ime servera\ime_deljenog_resursa_ime_fajla (za server; detalji o imenima se dobijaju od administratora mreže).
Ako, pak, želimo da otvorimo neku Web stranicu koja se nalazi na serverima Interneta u polje Address: unosimo **URL** adresu u formatu:

protokol://lokacija putanja_do resursa

Primeri:
http://premier.net/~cogito/neuropsy.html (za prezentaciju iz neuropsihologije);

http://www.stat.ucla.edu/textbook/ (za prezentaciju o osnovama statističke analize);

http://www.yahoo.com (za pozivanje poznatog pretraživača pomoću kojeg možemo lakše krstariti po okeanima Interneta)

Pri unošenju adrese deo koji se odnosi na protokol se može i izostaviti.
Kada se unese URL adresa i pritisne dugme **Enter** Explorer će posle izvesnog vremena "dovući" prezentaciju u svoj okvir. Za vreme dovlačenja prezentacije pokazivač miša ima izgled peščanog sata. Dokumenti u prezentaciji mogu imati hiperlinkove, najčešće u obliku podvučenog ili obojenog teksta ili sličice. Kada se pokazivač miša usmeri na link dobija oblik šake sa ispruženim kažiprstom. Klik na link levim dugmetom dovodi do izvršenja procedure koja je tim linkom predviđena.
Klikom na **desno** dugme miša po usmeravanju pokazivača na objekat u prezentaciji možemo dobiti meni mogućih akcija u vezi sa tim objektom. Tako, na primer, ako nanišanimo sliku koja je deo prezentacije i kliknemo na desno dugme možemo u meniju koji se pojavi kliknuti na **Save image as** i zapisati sliku na disk na koji želimo.
Za vraćanje na prethodno posećene strane unutar iste seanse treba kliknuti (levim dugmetom) na dugme **Back** a za odlazak u suprotnom smeru dugme **Forward** (⇒).
Dugme **Stop** služi za prekidanje započetog "dovlačenja" prezentacije ili fajla, a dugme **Refresh** za ponovno učitavanje prezentacije sa servera na kojem se nalazi. Ukoliko želimo da pogledamo prezentaciju koja se često menja (npr. najnovije vesti) može nam se desiti, ako pri tom ne kliknemo na dugme **Refresh** da Internet Explorer prikaže zastareli sadržaj koji je sačuvan privremeno u kešu (cash) pri prethodnoj poseti istoj stranici.
Dugme **Home** služi za vracanje na naslovnu stranicu, tj. onu stranicu koju vidimo prvu pri aktiviranju Internet Explorera.

Preuzimanje fajlova sa FTP sajtova

FTP serveri koriste File Transfer Protocol a za povezivanje na FTP sajt treba u polju **Address**: programa Internet Explorer uneti ime takvog sajta.
Prethodno treba znati da li FTP server na koji se vezujemo zahteva naše korisničko ime i lozinku jer tada u URL adresu moramo uneti i korisničko ime i lozinku.
Pri povezivanju na određeni FTP sajt treba kliknuti na link fajla koji želimo da preuzmemo. U prozoru za dijalog **File Download**, koji se pojavljuje u tom trenutku, treba odabrati opciju **Save this program (file)**, ako već nije uključena, i potom kliknuti na dugme **OK**. Na kraju, treba specifikovati disk i folder u koji će fajl biti snimljen.

DODATAK 2: OSNOVE TEORIJE VEROVATNOĆE I MATEMATIČKE STATISTIKE NEOPHODNE ZA RAZUMEVANJE STATISTIČKOG ZAKLJUČIVANJA

SADRŽAJ DODATKA 2

I. **OSNOVNI POJMOVI I PRINCIPI TEORIJE VEROVATNOĆE** 2

I.1. Osnovni pojmovi 2
I.2. Osnovni pojmovi i principi kombinatorike 3
I.3. Matematička definicija verovatnoće 5
I.4. Interpretacije verovatnoće 6
I.5. Uslovna verovatnoća 8
I.6. Osnovne teoreme o verovatnoći 8

II. **SLUČAJNE VARIJABLE** 10

II.1. Pojam slučajne varijable 10
II.2. Distribucije verovatnoća, funkcije gustine i funkcije distribucije slučajnih varijabli 11
II.3. Distribucija uzorkovanja (sampling distribution) 13
II.4. Opis distribucije populacije(očekivana vrednost, varijansa i momenti distribucije) 14
II.5. Prikaz odabranih distribucija verovatnoća i funkcija gustine 16
Binomna distribucija 16
Normalna ili Gaus - Bernulijeva funkcija gustine 19
Hi - kvadrat funkcija gustine 21
Studentova t funkcija gustine 22
Snidikorova f funkcija gustine 23

III. **POVEZANOST TEORIJSKIH DISTRIBUCIJA VEROVATNOĆA I EMPIRIJSKIH ISTRAŽIVANJA POJAVA** 24

III.1. Odnos matematičkog modela i prirode pojave na koju se on primenjuje 25
III.2. Odnos teorijskih distribucija verovatnoća i empirijskih distribucija učestalosti 25

I. OSNOVNI POJMOVI I PRINCIPI TEORIJE VEROVATNOĆE

Primer: Da li je moguće preneti drugom čoveku (primaocu) informacije o opaženom objektu bez posredovanja primaočevih čula? Postoji li Ekstra Senzorna Percepcija? Kako se može odgovoriti na ovo pitanje? Jedan od mogućih načina je i ovaj:
Pet raznih simbola na kartama (valjano promešanim) ponavljaju se tako da svaki simbol bude ponovljen 5 puta u seriji od 25 karata. Odašiljač izvlači kartu po kartu, čita je jedva čujno, a primalac treba da kaže koja je karta izvučena. Karta se vraća u špil pa se zatim izvlači sledeća karta za prenos. Rezultat: U seriji od 25 pokušaja pogođeno je 8 karata. Da li rezultat pruža osnovu za pretpostavku o postojanju ESP? Ili je u pitanju čist slučaj? Možemo naći matematički model po kojem ćemo znati koliko tačnih odgovora očekujemo na osnovu delovanja samo 'zakona slučaja'. Da li 8 tačnih odgovora može biti posledica slučaja i, ako može, kolika je verovatnoća njegovog javljanja samo na osnovu slučaja? Ako se samo na osnovu slučaja ovaj rezultat može pojaviti veoma retko kažemo da on govori u prilog postojanju ESP. Dakle, odgovor na ovo pitanje koje zanima mnoge psihologe nije moguć bez poznavanja osnovnih principa i računa verovatnoće.

I. 1. Osnovni pojmovi

Eksperiment

Dobro definisan čin ili proces koji vodi jednom dobro definisanom ishodu. (Winkler, str..42) Kada se inicira neki dinamički proces koji generiše neki događaj i posmatraju rezultati takvog procesa, to je eksperiment (Stilson, str.42).

Slučajni eksperiment

Slučajni ili statistički eksperiment je praćenje neke aktivnosti ili proces u kome se vrši određeno merenje. Drugačije rečeno, to je svaki preciziran (planiran) proces posmatranja ili prikupljanja podataka koji se može **ponavljati neograničen**, tj. proizvoljan **broj puta** u **istim uslovima**. Slučajni eksperiment ima brojne **ishode**, a pojedinačna pojava određenog ishoda se ne može unapred predvideti sa potpunom izvesnošću. Svako posmatranje ili prikupljanje podataka, tj. svaki statistički eksperiment ima samo jedan ishod iz skupa mnogih mogućih ishoda koji su međusobno isključivi.
Primeri statističkih eksperimenata: bacanje novčića, bacanje kocke, anketiranje ispitanika o nekoj pojavi, merenje vremena reakcije...

Pokušaj (engl. trial)

Pojedinačno izvođenje slučajnog eksperimenta (npr. jedno bacanje novčića).

Ishod (engl. outcome)

Ishod je određeni, realizovani rezultat eksperimenta ili mogući rezultat jednog pokušaja. Na primer mogući ishodi u jednom bacanju novčića su P(ismo) i G(rb) dok su mogući ishodi u dva bacanja novčića PP, GG, GP i PG. Ishodi moraju da se međusobno isključuju, tj. da u jednom pokušaju bude moguć samo jedan ishod.

Prostor ishoda (engl. sample space) ili prostor elementarnih događaja nekog eksperimenta je skup svih mogućih distinktivnih ishoda jednog eksperimenta. Osnovni principi kombinatorike koji su neophodni za računanje mogućih ishoda eksperimenta dati su u odeljku **I. 2.** ovog dodatka.

Događaj (engl. event)

Događaj je bilo koji skup ishoda, tj. bilo koji podskup prostora ishoda. Događaj je jedan ili više ishoda, realizacija eksperimenta. Može biti prost (samo jedan izlaz eksperimenta, npr.realizacija jedne bilo koje strane kocke) ili složen (obuhvata više od jedne realizacije eksperimenta, npr.pojava neparnog broja kocke).

Slučajan događaj (engl. random event) je onaj za koji pod svim poznatim uslovima prilikom izvođenja nekog eksperimenta nije moguće predvideti da li će se (ili neće) desiti.

Verovatnoća (engl. probability) je mera veličine mogućnosti ostvarenja ili neostvarenja događaja. Do nje se dolazi apriornom logičkom analizom procesa koji vodi određenom ishodu, pretpostavkama ili empirijskom procenom. Iskazuje se razlomcima (1/10), proporcijama (0.28) ili procentima (33%).

I. 2. Osnovni pojmovi i principi kombinatorike

(U ovom delu teksta će za operaciju množenja, kada je označavanje operacije neophodno biti korišćen znak *).

Da bi se izračunale verovatnoće pojedinih ishoda potrebno je odrediti sve moguće ishode eksperimenta. Kada je broj mogućih ishoda mali lako je odrediti koliko ishoda imamo i koji su to ishodi. Nekada to nije baš lako pa se koriste određena pravila:

1. Osnovno pravilo kombinatorike: Ako su rezultati eksperimenta k događaja, od kojih se prvi može desiti na n_1 načina, drugi na n_2 načina i tako redom, onda se svih k događaja mogu desiti na $n_1 * n_2 * \ldots * n_k$ načina.

2. Permutacije predstavljaju svaki mogući redosled, poredak n elemenata. Broj mogućih permutacija od n elemenata jednak je n faktorijel (u oznaci n!):

   nPn = n!/(n-n)! = n! jer je faktorijel 0 (u oznaci 0!) jednak 1.

> **Zapamtite:** n faktorijel (n!) je proizvod svih brojeva od n do 1.
>
> Na primer: 3! je 3*2*1=6.

Primer permutacija: Na koliko sve načina se mogu poređati 4 različita broja?
Budući da je 4! jednako 4*3*2*1=24, 4 različita broja se mogu poređati na 24 načina.

3. Permutacije s ponavljanjem Ako među n elemenata ima n_1 međusobno jednakih , n_2 međusobno jednakih..., n_r međusobno jednakih, a pri tome je $n_1 + n_2 + \ldots + n_r = \sum_{1}^{r} n_i = n$, onda je broj načina na koji se n objekata može grupisati u r klasa sa n_i elemenata u

svakoj klasi jednak:

$$nPn_1, n_2 \ldots n_r = n!/(n_1!n_2! \cdots n_r!) = \binom{n}{n_1\, n_2 \ldots n_r}.$$

Prethodni izraz $\binom{n}{n_1\, n_2 \ldots n_r}$ čita se kao n **nad** n_1 , n_2 ... n_r i zove se **multinomijski koeficijent**, a pojavljuje pri razvoju izraza

$$(x_1 + x_2 + \ldots + x_r)^n = \sum \binom{n}{n_1\, n_2 \ldots n_r} x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r},$$

pri čemu se sumacija vrši za sve nenegativne cele brojeve $n_1, n_2, \ldots, n_r$ tako da je $n_1 + n_2 + \ldots + n_r = n$. Poznavanje ovog koeficijenta može pomoći u razumevanju multinomijske raspodele (odeljak **II.5.**)

4. Varijacije predstavljaju broj permutacija za samo r elemenata od svih n:

$$nPr = n\,(n-1)\ldots(n-r+1) = n!\,/\,(n-r)!$$

Ovaj obrazac omogućuje da izračunamo i broj uzoraka veličine r koji se mogu izvući iz populacije veličine n bez vraćanja, ali samo ako uzorak shvatimo kao **uređen** skup, tj. ako različiti redosledi istih elemenata predstavljaju različite uzorke.

5. Varijacije s ponavljanjem:

$$nPr = n^r$$

Ovaj obrazac omogućuje da izračunamo i broj uzoraka veličine r koji se mogu izvući iz populacije veličine n sa vraćanjem. Isto tako, pomoću ovog obrasca možemo izračunati broj mogućih načina da se odgovori na testu koji ima r pitanja, a na svako pitanje je ponuđeno n odgovora. Tako, na primer, ako ima 4 pitanja, a ponuđeni su samo odgovori DA ili NE, broj mogućih složajeva odgovora je $2^4 = 16$.

6. Kombinacije predstavljaju mogući broj grupisanja r elemenata od ukupno n elemenata, ali tako da se nizovi međusobno razlikuju bar po jednom elementu (različiti redosledi istih elemenata predstavljaju jedan isti niz).

$$_nC_r = {}_nP_r\,/\,r!$$

ili

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!}.$$

Pri tome, važe i sledeće jednakosti: $_nC_r = {}_nC_{n-r} = {}_{n-1}C_r + {}_{n-1}C_{r-1}$.

Izraz $\binom{n}{r}$ je binomijalni koeficijent, a čita se kao n **nad** r. Ovaj koeficijent se pojavljuje u razvoju binomnog izraza $(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$ i koristi se u računanju verovatnoće kod binomijalne raspodele (odeljak **II. 5.** ovog dodatka).

Primer: Koliko mogućih kombinacija od po 4 broja možemo napraviti iz skupa od 6 brojeva?

Rešenje:

$$_6C_4 = 6! / (4!2!) = 15$$

7. Kombinacije sa ponavljanjem nam mogu pomoći da odredimo broj grupa veličine r koje se mogu dobiti iz populacije veličine n sa vraćanjem:

$$_nC_r = \binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$$

8. Hipergeometrijsko pravilo: Ako osnovni skup sadrži dve vrste elemenata : n_1 jedne i n_2 druge vrste, broj podskupova veličine r koje možemo formirati iz skupa n je:

$$n_1 Cr_1 * n_2 Cr_2$$

pri čemu je $n_1 + n_2 = n$, dok je $r_1 + r_2 = r$. Ovo pravilo korisno je za razumevanje hipergeometrijske raspodele (odeljak **II.5.** dodatka)

I. 3. Matematička definicija verovatnoće

Ako je dat prostor ishoda S i neki skup događaja A u S, funkcija verovatnoće pridružuje svakom događaju A u S, tj. svakom podskupu A iz S, realni broj P(A), verovatnoću događaja A, tako da su ispunjena, tj. važe tri aksioma:

1. **Nenegativnost**: $P(A) \geq 0$ za svako A (nema negativne površine na Venovom dijagramu[1])
2. **Normiranost**: $P(S) = 1$, tj. površina pravougaonika na Venovom dijagramu = 1
3. **Aditivnost**: Ako prebrojivi podskupovi događaja A_1, A_2,..., A_k postoje i ako su međusobno isključivi (nemaju zajedničkih elemenata) verovatnoća njihove unije jednaka je zbiru njihovih pojedinačnih verovatnoća:

$$P(A_1 \cup A_2 \cup ... \cup A_k) = P(A_1)+P(A_2)+...+P(A_k)$$

Dakle, realni brojevi koji se pridružuju svakom događaju, tj. podskupu prostora ishoda, su verovatnoće ako su:

1. nenegativni;

2. jednaki jedinici za ceo skup, tj. prostor ishoda, i

3. zbir brojeva dodeljenih isključivim događajima jednak broju dodeljenom njihovoj uniji.

---

[1] Venov (John Venn, matematičar) dijagram predstavlja prostor ishoda pravouganikom (ili krugom) a skupove događaja površinama unutar pravouganika ili kruga.

Matematička (Kolmogorovska) definicija verovatnoće (prema Ivković (1990), str.12):

Verovatnoća P(.) je numerička funkcija definisana nad σ - poljem događaja T koja ima sledeće osobine:[2]

1. nenegativnost: za svako $A \in T$ važi $P(A) \geq 0$
2. normiranost: $P(\Omega) = 1$, pri čemu je Ω skup svih logički mogućih ishoda nekog opita.
3. σ-aditivnost : Ako su $A_n \in T$, $n = 1,2,\ldots$, uzajamno isključujući događaji (tj. $A_i \cap A_j = \varnothing$ za $i \neq j$, onda je $P(\sum_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

I. 4. Interpretacije verovatnoće

1. Statistička: Verovatnoća kao relativna frekvencija događanja nekog događaja u velikom broju posmatranja, tj. ponavljanja eksperimenta ili kao granična vrednost relativne frekvence.

Verovatnoća je broj kome se približava f_A/n kako n teži beskonačnosti:

$$P(A) = \lim_{n \to \infty} \frac{f_A}{n}$$

Zakon velikih brojeva (prema Winkler, str. 66):

$$\text{Ako } n \to \infty, \text{ onda } P\left(\left|\frac{f_A}{n} - p\right|\right) > \varepsilon \to 0$$

Ako je p verovatnoća događaja A i ako se učini n pokušaja, nezavisno i pod istim uslovima, verovatnoća da će relativna frekvenca događaja A da se razlikuje od p više od nekog malog pozitivnog broja se približava nuli kako se broj pokušaja povećava.

Uočimo da ovo **ne znači da će sa povećanjem n da se smanjuje diskrepanca** |f/n - p|, već da sa **povećanjem n velika diskrepanca** postaje **malo verovatna**! Verovatnoća da će |f/n-p biti veće od malog pozitivnog broja opada!

---

[2] Za klasu T događaja A. B. C....koje posmatramo kod ogleda sa skupom ishoda Ω se zahteva da bude zatvorena u odnosu na operacije uzimanja komplementa (tj. suprotnog događaja koji se realizuje AKKO - ako i samo ako- se ne realizuje događaj čiji je to komplement) i prebrojive unije $(\bigcup_{i=1}^{\infty} A_k = \{w: \exists k, w \in A_k\}$

i preseka $\bigcap_{i=1}^{\infty} A_i = \{w: \forall k, w \in A_k\}$ gde je w pojedini ishod eksperimenta.

Aksiom: Klasa T događaja koji se posmatraju kod nekog opita sa slučajnim ishodima čini σ -polje (σ - algebru) tj.:

1. $\Omega \in T$;
2. Ako $A \in T$ tada $A^c \in T$, pri čemu je A^c suprotan događaj događaju A ;
3. Ako $A_n \in T$, n= 1,2,... tada $\bigcup_{n=1}^{\infty} A_n \in T$
4. $\varnothing \in T$;
5. Ako $A_n \in T$, n= 1,2,... tada $\bigcap_{n=1}^{\infty} A_n \in T$

Uređena trojka (Ω, T, P) je **prostor verovatnoća**.

U praksi, budući da se eksperiment može ponoviti samo konačan broj puta verovatnoću kao graničnu vrednost nemoguće je odrediti (čak i pod uslovom da ona postoji) pa se verovatnoća ocenjuje, ako je broj opservacija (pokušaja) veliki, preko relativne frekvence događaja:

$$p(A) = f_a/f_n$$

2. Klasična interpretacija verovatnoće - A priori verovatnoća

Ako se neki događaj A može desiti na a načina od ukupno mogućih n načina tada je verovatnoća da se dogodi događaj A (povoljan ishod) jednaka odnosu povoljnih ishoda i ukupno mogućih:

$$p(A)=a / n$$

a verovatnoća da se ne dogodi događaj A je:

$$q = (n-a) / n = b/n = 1- p.$$

Drugim rečima verovatnoća ovako definisana jednaka je odnosu broja povoljnih ishoda i broja svih mogućih ishoda. Verovatnoća, u skladu sa ovom koncepcijom, određuje se matematičkim putem, unapred, pre izvršenog eksperimenta. Ova se koncepcija verovatnoće zasniva na pretpostavci da su svi ishodi eksperimenta međusobno isključivi i jednako mogući.
Da bismo odredili verovatnoću a priori potrebno je znati broj ukupno mogućih ishoda n i broj povoljnih ishoda a.

$$n = a+b \; ; \; 1= a/n + b/n = p+q.$$

Odavde sledi da:

1. Ni p ni q ne mogu biti veći od 1;
2. Ni p ni q ne mogu biti negativni brojevi;
3. Apsolutna sigurnost dešavanja povoljnog ishoda predstavlja granicu verovatnoće, pri čemu je p=1 a q = 0;
4. Apsolutna neizvesnost događanja povoljnog ishoda predstavlja granicu verovatnoće, pri čemu je p = 0 a q =1;
5. p i q su odnosi;
6. p i q su komplementarni.

Primeri: Kolika je verovatnoća da će u bacanju novčića pasti 'grb'?

p= 1/2 =0.5

Kolika je verovatnoća da se iz dobro promešanog špila od 52 karte izvuče karta "pik"?

p=13/52 = 0.25

Verovatnoća rađanja muškog deteta je 0.515 a ne 0.50!

3. Subjektivna verovatnoća je verovatnoća za koju se neko opredeljuje na osnovu iskustva, uverenja i nedovoljno informacija. Ova je verovatnoća mera nečijeg verovanja da će se

događaj desiti. Ona ne podleže matematičkoj obradi.Koristi se onda kada ima malo ili nimalo informacija o sdašnjim događajima na osnovu kojih se može izračunati verovatnoća. Na primer: ekonomski eksperti procenjuju verovatnoću da inflacija bude četvorocifrena!

I. 5. Uslovna verovatnoća

Uslovna verovatnoća je mera mogućnosti, izvesnosti dešavanja događaja B pošto se prethodno desio događaj A. Prethodno dešavanje događaja A je uslov koji treba da bude ispunjen i pod kojim ispitujemo dešavanje događaja B.
Uslovna verovatnoća događaja B, pod uslovom da se desio događaj A, izračunava se po formuli:

$$P(B/A) = P(A \cap B) / P(A)$$

gde je $P(A \cap B)$ zajednička verovatnoća događaja A i B, a P(A) marginalna verovatnoća događaja A.

Zapamtite: Ako je P(B/A) = P(B) onda su A i B statistički nezavisni događaji.

I. 6. Osnovne teoreme o verovatnoći

Aditivna teorema

Verovatnoća da se desi jedan od dva događaja (ili događaj A ili događaj B) jednaka je zbiru verovatnoća individualnih događaja umanjenom za verovatnoću zajedničkog ostvarenja oba događaja:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Ako su događaji **međusobno isključivi** teorema dobija oblik:

$$P(A \cup B) = P(A) + P(B)$$

jer je u tom slučaju $P(A \cap B)$ jednako nuli.

Multiplikaciona teorema

Verovatnoća zajedničkog dešavanja dva događaja (A i B) jednaka je proizvodu verovatnoće događaja A i uslovne verovatnoće događaja B:

$$P(A \cap B)=P(A) * P(B/A).$$

Ako su događaji međusobno nezavisni onda je :

$$P(A \cap B) = P(A) * P(B)$$

jer je u tom slučaju uslovna verovatnoća P(B/A) jednaka P(B).

Bajesovo (Bayes) pravilo

Ako su A i $B_1, B_2, \ldots, B_n$ događaji u skupu ishoda S, pri čemu su B_i međusobno isključivi, a njihova unija jednaka prostoru ishoda, dakle:

a) $\bigcup_{i=1}^{n} B_i = S$;

b) $(\forall i).\ P(B_i) > 0$;

c) $B_i \cap B_j = 0$, za $i \neq j$,

tada je

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_j)}.$$

Bajesovo pravilo može, na primer, poslužiti za određivanje verovatnoće da osoba ima određenu bolest ako su poznate apriorna verovatnoća oboljevanja od te bolesti, $P(B_j)$, i verovatnoća da osoba ima određeni simptom (A) ako ima neku bolest B_j, tj. $P(A|\ B_j)$.
U statistici, u okviru tzv. Bajesovskog pristupa, koji nije obrađen u ovom Priručniku, ovo pravilo služi kao osnova za određivanje aposteriorne verovatnoće neke hipoteze s obzirom na dobijene podatke, ako je poznata njena apriorna verovatnoća.

Formula potpune verovatnoće:

Ako su $B_1, B_2, \ldots, B_n$ događaji u skupu ishoda S, pri čemu su B_i međusobno isključivi, verovatnoća svakog od njih veća od nule, a njihova unija jednaka prostoru ishoda, dakle:

a) $\bigcup_{i=1}^{n} B_i = S$;

b) $(\forall i),\ P(B_i) > 0$;

c) $B_i \cap B_j = 0$, za $i \neq j$,

tada za bilo koji događaj A važi:

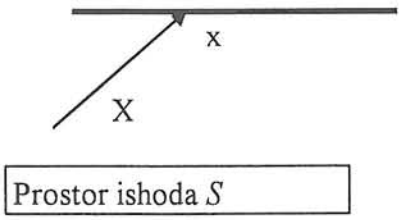
$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Formula potpune verovatnoće može korisno poslužiti ako nije sasvim jasno kako bismo izračunali verovatnoću nekog događaja A, a na raspolaganju su nam uslovne verovatnoće $P(A|\ B_i)$ i verovatnoće događaja B_i, tj. $P(B_i)$.

II. SLUČAJNE VARIJABLE

II. 1. Pojam slučajne varijable

Izvođenje nekog posmatranja u empirijskom istraživanju može se posmatrati kao izvođenje slučajnog eksperimenta jer rezultat, ishod opservacije, nije unapred poznat.
Primer: Ukoliko se zadaje neki merni instrument za ispitivanje agresivnosti svaki složaj odgovora na pitanja iz instrumenta se može posmatrati kao ishod u prostoru uzorka slučajnog eksperimenta "zadavanje skale agresivnosti slučajno izabranom pojedincu". Svakom se složaju prema pravilima pridružuje određeni skor (ili mera) koji predstavlja vrednost na nekoj varijabli. Pravilo za dodeljivanje određenog skora određenom složaju odgovora na stavkama skale je slučajna varijabla, a proces zadavanja skale i primene ovog pravila je posmatranje vrednosti slučajne varijable (cf. Stilson, str.122). Pravilo (funkcija) koje dodeljuje jednu vrednost (realni broj) svakom ishodu u prostoru uzorka je slučajna varijabla koja je definisana na prostoru uzorka S.
Dakle slučajna varijabla bi prema ovom shvatanju bila funkcija koja svakom ishodu (elementarnom događaju) prostora ishoda S dodeljuje jednu (i samo jednu) numeričku vrednost:



Na crtežu X je **slučajna varijabla**, a x je **vrednost** slučajne varijable.
Slučajna varijabla se može definisati i kao veličina koja podrazumeva jednu i samo jednu brojnu vrednost kao ishod jednog eksperimenta ili kao numerička funkcija koja svakom ishodu statističkog eksperimenta pridružuje jedan realan broj.Jedna vrednost slučajne varijable može biti povezana sa više elementarnih ishoda eksperimenta ali svakom elementarnom ishodu odgovara samo jedna vrednost slučajne varijable.
Na primer, broj koji se pojavljuje prilikom bacanja kocke je slučajna varijabla koja može imati vrednosti od 1 do 6 ali kao rezultat, ishod jednog bacanja može se javiti samo jedan broj.
Prema B. Ivanoviću (1966) slučajna promenljiva je promenljiva X koja na slučaj može uzimati jednu od vrednosti $x_1, x_2,..,x_n$ sa odgovarajućim verovatnoćama $p_1,p_2,..,p_n$ tako da je

$$\sum_{i=1}^{n} p_i = p_1 + p_2 +...+ p_n = 1.$$

Prema Demingu (Deming, str.83) slučajna varijabla je rezultat primene slučajne operacije. Ponavljanje slučajne operacije daje niz vrednosti slučajne varijable čija se granična (teorijska) ili distribucija verovatnoća može matematički izvesti.Ako slučajnu operaciju ponovimo n puta dobijamo n vrednosti slučajne varijable ili slučajni uzorak veličine n.

Veoma važno: **Ako je X slučajna varijabla svaka funkcija od X je takođe slučajna varijabla**.[3] Dakle, budući da su aritmetička sredina, suma, medijana ili bilo koji drugi statistik funkcija slučajne varijable oni su takođe slučajne varijable.

Diskretne slučajne varijable

Mogu uzeti bilo konačan ili prebrojivo beskonačan broj vrednosti.[4]
Primeri diskretnih slučajnih varijabli: broj grbova u 5 bacanja novčića:0,1,2,3,4,5; broj dece u porodici; broj grešaka u eksperimentu.

Kontinuirane slučajne varijable

Mogu teorijski uzeti neprebrojivo beskonačan broj vrednosti. Između bilo koje dve vrednosti takve varijable je neodređeno veliki broj vrednosti.[5]

Primeri: inteligencija; visina; autoritarnost; brzina reagovanja.
Ipak, realno (u istraživanju) mi kontinuiranu varijablu pretvaramo u diskretnu a u statističkoj analizi podatke tretiramo (radi lakšeg rada) kao da su dobijeni na kontinuiranoj varijabli.

II. 2. Distribucije verovatnoća, funkcije gustine i funkcije distribucije slučajnih varijabli

Distribucija verovatnoća diskretne slučajne varijable

Je skup parova koje čine moguće vrednosti slučajne varijable i svakoj vrednosti pridružene verovatnoće da varijabla uzme datu vrednost

X	p
$X = x_1$	$P(X = x_1)$
$X = x_2$	$P(X = x_2)$
----------	----------
$X = x_n$	$P(X = x_n)$
	$\sum_{i=1}^{n} P(X = x_i) = 1$

[3] Funkcija (funkcijska relacija) je relacija u kojoj je svaki član domena sparen sa jednim i samo jednim članom kodomena. Matematička relacija je podskup Dekartovog proizvoda. tj. specifičan skup parova od svih mogućih parova dva skupa. Može se specifikovati navođenjem svih parova ili (češće) iskazom o pravilu koje opisuje parove : $S=\{(a,b) \in A \otimes B) \mid$ a je oženjen b}, pri čemu je A skup svih muškaraca a B skup svih žena.

[4] Prebrojivo beskonačan znači da se članovi skupa mogu postaviti u odnos 1:1 sa pozitivnim celim brojevima tako da je svaki član povezan sa jednim i samo sa jednim pozitivnim celim brojem.

[5] Neprebrojivo beskonačan skup je onaj čiji su elementi tako "gusti" da primena pozitivnih brojeva ne može da ih iscrpe.

Primer: Distribucija verovatnoća za broj "pisama" u dva bacanja novčića:

X	p
X = 0	1/4
X = 1	1/2
X = 2	1/4

Funkcija distribucije ili kumulativna funkcija distribucije (u daljem tekstu KFD) za diskretnu slučajnu varijablu daje verovatnooću da slučajna varijabla uzme vrednost manju ili jednaku nekoj određenoj vrednosti:

$$F(x_a) = P(X \leq x_a) = \sum_{i=1}^{n} P(X = x_i)$$

Primer: KFD za broj "pisama" u dva bacanja novčića:

$X \leq 1$	3/4
$X \leq 2$	1

Funkcija gustine verovatnoće kontinuirane slučajne varijable

Za kontinuiranu slučajnu varijablu umesto o verovatnoći da ona uzme određenu vrednost (ova je verovatnoća u stvari jednaka nuli!) potrebno je odrediti verovatnoću da slučajna varijabla uzme neku vrednost u intervalu od .. do. Umesto o verovatnoći da varijabla uzme neku određenu vrednost x_i govorimo o gustini verovatnoće varijable X na x_i /u oznaci $f(x_i)$/.[6]

Dakle, ako je X kontinuirana slučajna varijabla i ako je f(x) funkcija od X tako da za bilo koje dve vrednosti X, a i b (tako da je a < b)

$$P(a < X < b) = \int_a^b f(x)\,dx = \lim_{k \to \infty} \left[\sum_{i=1}^{k} f(x_i)\, h \right]$$

f(x) je funkcija gustine verovatnoće za X.[7]

Dakle, $P(a \leq X \leq b) = P(a < X < b)$. Jedna tačka ima širinu 0 pa je površina iznad nje 0.
Zbog grubosti mernih instrumenata obično teorijski kontinuirana slučajna varijabla u istraživanju ima diskontinuirane vrednosti.

---

[6] Gustina verovatnoće:. Verovatnoća da slučajna varijabla uzme vrednost u intervalu od granica a do b: $P(a \leq X \leq b) = P(X \leq b) - P(X < a)$. Ako veličinu intervala b - a označimo sa Δx tada je $b = a + \Delta x$. Verovatnoća intervala u odnosu na Δx je: $\frac{P(a \leq X \leq b}{\Delta x} = \frac{P(X \leq a + \Delta x) - P(X < a)}{\Delta x}$. Ako a fiksiramo dok Δx varira i recimo približava se 0, tada se ceo količnik menja i približava određenoj graničnoj vrednosti kako Δx ide bliže 0. Ova granica daje gustinu verovatnoće varijable X na vrednosti a. Grubo rečeno, gustina verovatnoće na a za najmanju promenu veličine intervala daje brzinu promene verovatnoće intervala sa gornjom granicom a. (cf. Winkler, p.129).

[7] Izraz f(x) dx je površina pravougaonika čija je visina f (x) a osnovica razmak h = (x - dx/2, x + dx/2).

Funkcija distribucije ili kumulativna funkcija gustine (u daljem tekstu KFG za kontinuiranu slučajnu varijablu

Ako je f(x) funkcija gustine verovatnoće kontinuirane slučajne varijable KFG :

$$F(x_a) = P(X < x_a) = \int_{-\infty}^{x_a} f(x)\,dx \text{ za svako } x_a$$

daje verovatnoću da na slučaj uzeta vrednost varijable padne između -∞ (ili x_{min}) i x_a. Verovatnoća je dakle jednaka površini ispod krive koja predstavlja funkciju gustine ili visini KFG.

Razlika diskretnih i kontinuiranih slučajnih varijabli u svetlu funkcije distribucije

Slučajna varijabla je diskretna ako je njena funkcija distribucije stepeničasta a kontinuirana ako je njena FD kontinuirana funkcija.

Odnos KFG i funkcije gustine za kontinuiranu slučajnu varijablu

$$f(x) = F'(x)$$

Dakle, prvi izvod KFG jednak je funkciji gustine verovatnoće.

*****

Distribucija verovatnoća i funkcije gustine odnose se na populaciju.[8] Na osnovu njih možemo odrediti verovatnoću da neki član populacije izvučen na slučaj ima određenu vrednost (za diskretne varijable) ili meru u nekom intervalu između datih granica (za kontinuirane slučajne varijable). Budući da se gotovo nikada ne ispituje cela populacija (a često je to i nemoguće) do ditribucija ili funkcija gustine verovatnoće se dolazi:
1. na osnovu analize logičke strukture, tj. na osnovu prirode procesa koji dovodi do opservacije ili
2.ocenjivanjem na osnovu rezultata dobijenih na uzorku tj. na osnovu distribucije učestalosti.

II. 3. Distribucija uzorkovanja (sampling distribution)

Za razliku od **parametara** koji su **konstante**, svaki statistik (numerička karakteristika uzorka) se pre uzimanja uzorka posmatra kao slučajna varijabla. **Statistik (mera uzorka) je ocenjivač parametra (mere populacije)**. Zaključci o parametrima populacije na osnovu numeričkih svojstava uzorka su uvek podložni grešci čija veličina može biti određena. **Teorijska distribucija verovatnoća (funkcija gustine) koja pokazuje funkcionalnu vezu mogućih vrednosti statistika (statističke mere uzorka) i odgovarajućih verovatnoća da statistik uzme neku vrednost (ili neku vrednost u određenom intervalu) za sve moguće uzorke određene veličine je distribucija uzorkovanja (uzoračka ili sampling distribucija)**.
Distribucija uzorkovanja predstavlja teorijsku distribuciju koja predstavlja sve moguće vrednosti koje bi statističke mere uzoraka mogle uzeti i njima pridružene relativne frekvence (verovatnoće) sa kojom ove vrednosti mogu da se jave ako bismo uzeli veoma veliki (neograničeni) broj merenja na reprezentativnim uzorcima veličine n iz određene definisane

[8] Prema Stilson-u (str.31) populacija se može shvatiti i kao skup entiteta na koje želimo da primenimo zaključke na osnovu rezultata istraživanja dobijene na uzorku, tj podskupu datog skupa

populacije. Tretiranje statistika uzoraka kao slučajnih varijabli počiva na proceduri uzorkovanja - odabiranje uzorka vršimo na osnovu slučaja. Distribucije uzorkovanja dobijene su najčešće matematičkim izvođenjima zasnovanim na pretpostavkama o tome kako deluju slučajni činioci. Onda kada je njihovo izvođenje vrlo složeno, ili nemoguće, njihova aproksimacija se dobija tzv. **bootstrap** postupcima, tj. simuliranjem izvlačenja velikog broja slučajnih uzoraka i pravljenjem distribucije statistika pod datim uslovima.
Distribucija uzorkovanja nije isto što i distribucija populacije (sem ako je n =1), ali od **nje zavisi** na specifičan način.

II. 4. Opis distribucije populacije (očekivana vrednost, varijansa i momenti distribucije)

Očekivana vrednost slučajne varijable

Za diskretne slučajne varijable:

$$E(X)= \Sigma P(X= x_i)\, x_i = \Sigma p_i\, x_i = \mu$$

Dakle, očekivana vrednost je unutrašnji proizvod (tj.suma proizvoda) mogućih vrednosti varijable i njima pripadajućih verovatnoća.

Za kontinuirane slučajne varijable:

$$E(X) = \int x\, f(x)\, dx$$

Očekivana vrednost slučajne varijable je aritmetička sredina distribucije verovatnoća slučajne varijable u populaciji.

Osobine očekivane vrednosti:
1. $E(a) = a$, pri čemu je a konstanta;
2. $E(aX) = a\,E(X)$, pri čemu je a konstanta a X slučajna varijabla;
3. $E(X+ a) = E(X) + a$
4. $E(X+Y) = E(X) + E(Y)$, pri čemu su X i Y slučajne varijable;
5. $E(a + b_1X_1 + b_2X_2 +\ldots+ b_jX_j) = a + b_1\,E(X_1) + b_2\,E(X_2) +\ldots+ b_j\,E(X)_j$. pri čemu su a i b-ovi konstante dok je izraz $a + b_1X_1 + b_2X_2 +\ldots+ b_jX_j$ linearna kombinacija slučajnih varijabli.
6. $E(XY) = E(X) * E(Y) + \rho_{XY}\,\sigma_X\,\sigma_Y$. pri čemu je ρ_{XY} Pirsonov koeficijent korelacije.
Očigledno, ako je $\rho_{XY} = 0$, onda je $E(XY) = E(X) * E(Y)$.

Varijansa populacije σ^2

Jednaka je očekivanju kvadriranih odstupanja od aritmetičke sredine populacije:

$$\sigma^2 = E\,(X - \mu)^2 = E\,[X- E\,(X)]^2 = E\,(X^2) - [E(X)]^2 \Rightarrow E\,(X^2) = \sigma^2 + [E(X)]^2$$

Varijansa zbira slučajnih varijabli:

$\sigma^2_{X1+X2+....+Xm} = \sigma^2_{X1} + \sigma^2_{X2} + ... + \sigma^2_{Xm} + 2\,\sigma_{X1X2} + 2\sigma_{X2X3} + 2\sigma_{Xm-1\,Xm}$

gde su $\sigma^2_{X1}, \sigma^2_{X2}, ... \sigma^2_{Xm}$ varijanse $X_1, X_2, ..., X_m$ slučajnih varijabli a σ_{X1X2} , σ_{X2X3} , $\sigma_{Xm-1\,Xm}$ su m(m-1)/2 kovarijansi parova varijabli.

Varijansa razlike dveju slučajnih varijabli:
$\sigma^2_{X1-X2} = \sigma^2_{X1} + \sigma^2_{X2} - 2\,\sigma_{X1X2}$
gde su σ^2_{X1} i σ^2_{X2} varijanse X_1, X_2 slučajnih varijabli a σ_{X1X2} je kovarijansa varijabli.

Očigledno, ako su varijable uzajamno nezavisne, varijansa njihovog zbira(razlike) jednaka je zbiru (tj.razlici) varijansi pojedinačnih varijabli koje ulaze u zbir (koje čine razliku).

Varijansa linearne kombinacije slučajnih varijabli:

Ako su A, B_1, B_2,...,B_m konstante a X_1, X_2,...,X_m slučajne varijable varijansa linerane kombinacije slučajnih varijabli je:

$\sigma^2_{A-B1X1-...-BmXm} = B_1^2\sigma^2_{X1} + ... + B_m^2\sigma^2_{Xm} + 2\,B_1\,B_2\,\sigma_{X1X2} + ... + 2\,B_{m-1}\,B_m\,\sigma^2_{Xm-1\,Xm}.$

Standardna devijacija populacije:
$\sigma = \sqrt{\sigma^2}$

Pored očekivane vrednosti (koja opisuje centralnu tendenciju) i varijanse (koja opisuje raspršenje) korisno je navesti i momente distribucije.

Momenti su očekivanja različitig stepena oko neke vrednosti.
Očekivanja oko nule, tj.početka slučajne varijable su:
E(X) je aritmetička sredina, $E(X^2)$...

Očekivanja oko aritmetičke sredine su centralni momenti:

$$E[(X-\mu)^r]$$

$E[(X-\mu)^1]$ je prvi centralni momenat

$E[(X-\mu)^2]$ je drugi centralni momenat

$E[(X-\mu)^3]$ je treći centralni momenat

$E[(X-\mu)^4]$ je četvrti centralni momenat

$\frac{E[(X-\mu)^3]}{\sigma^3}$ je Skewness (skjunis) ili pokazatelj asimetrije

$\frac{E[(X-\mu)^4]}{\sigma^4} - 3$ je Kurtosis (kurtozis) ili pokazatelj izduženosti

II. 5. Prikaz odabranih distribucija verovatnoća i funkcija gustine

Teorijske distribucije verovatnoća, odnosno, modeli ili zakoni verovatnoće definišu vezu između vrednosti diskretne slučajne varijable i verovatnoća sa kojima varijabla može uzeti bilo koju od tih vrednosti.
Najpoznatiji modeli distribucija za diskretne slučajne varijable su:
-binomni;
-Puasonov;
-uniformni;
-hipergeometrijski.

Uniformna distribucija

Formula uniformnog zakona verovatnoće je:

$$P(X= x)=1/n$$

pri čemu je P(X= x) verovatnoća sa kojom varijabla uzima vrednost x a n broj mogućih vrednosti koje može da uzme varijabla .

Zapamtite: Diskretna rektangularna tj.uniformna distribucija je korisna za rad sa rangovima: Ako su vrednosti slučajne varijable prvih r pozitivnih brojeva i ako je P(X=x)=1/r onda je
E(X)= (r+1)/2 a varijansa:$(r^2-1)/12$[9]

Binomna distribucija

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

gde je $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ binomni koeficijent a n i p su parametri.

Daje verovatnoću da će x elemenata slučajnog uzorka veličine n izvučenog sa vraćanjem iz binomne populacije ili bez vraćanja iz beskonačne binomne populacije biti iz poskupa A populacije.[10] Drugačije rečeno ona daje verovatnoću x "uspeha" u n pokušaja Bernulijevog procesa.[11]

---

[9] Za izvođenje ovih obrazaca poslužila su sledeća pravila: Zbir prvih r pozitivnih celih brojeva je $\sum_{x=1}^{r} x = \frac{r(r+1)}{2}$ a zbir njihovih kvadrata $\sum_{x=1}^{r} x^2 = \frac{r(r+1)(2r+1)}{6}$

[10] Binomna populacija S je sastavljena od dva međusobno isključiva i iscrpna podskupa A i $S \sim A$ (S minus A) tako da je verovatnoća da na slučaj izvučen član populacije bude iz podskupa A jednaka p a da bude iz podskupa $S \sim A$ jednaka q = 1-p.

[11] Bernulijev proces se sastoji od niza nezavisnih, identičnih Bernuli pokušaja (moguća su samo dva ishoda: "uspeh" i "neuspeh") gde p (verovatnoća "uspeha" u svakom pojedinom pokušaju) ostaje konstantno iz pokušaja u pokušaj.

Ako su p i q komplementarni tj. $p + q = 1$

$(p + q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x}$ gde x ide od 0 do n a n-x naravno od n do 0.

Očekivana vrednost varijable koja ima binomnu distribuciju verovatnoća je $\mu = np$ dok je varijansa $\sigma^2 = np(1-p)$.

Koeficijent asimetrije je: $(q - p)^2 / npq$ a koeficijent izduženosti: $(1-6pq/ npq)$.

Binomna raspodela je simetrična kada je $p = q = 0.5$ a što se p i q više razlikuju to je nesimetričnija. Kada $n \rightarrow \infty$ binomna raspodela teži normalnoj.

Primer: Ako je verovatnoća da je student psihologije muško 0.20, kolika je verovatnoća da će među 10 nasumice izvučenih studenata (sa vraćanjem) biti 6 muškaraca?

$P(X=6) = \binom{10}{6} 0.2^6 * 0.8^4 = 0.006$

Ako se 2/3 pacijenata sa dijagnozom shizofrenije na prvom pregledu oporavlja bez ikakvog tretmana da li se psihijatar koji je od 10 novoprimljenih pacijenata kod 8 zaključio da se oporavljaju može pohvaliti svojim terapijskim uspehom? Ne baš. Verovatnoća da se bar 8 od 10 njegovih poslednjih pacijenata spontano oporavi je 0.299.

$P(X \geq 8) = \sum_{x=8}^{10} \binom{n}{x} p^x q^{n-x} = 0.299$

Postoje gotove statističke tabele u kojima su date binomne verovatnoće za određene vrednosti n i p, a u programu SPSS ove verovatnoće se mogu izračunati u proceduri **Transform / Compute**, korišćenjem funkcije **CDF.binom (q, n, p)**, pri čemu su argumenti funkcije, tj. parametri distribucije: broj pokušaja (n), broj "uspeha" u n pokušaja manji od q, a p je verovatnoća "uspeha" u svakom pokušaju. Pri tom treba voditi računa da ova funkcija daje **kumulativnu** verovatnoću da u n pokušaja bude **manje od q** "uspeha".

Paskalova distribucija

Distribucija broja Bernulijevih pokušaja dok se ne dostigne x "uspeha" tj.daje verovatnoću da se u n pokušaja postigne unapred zadanih x "uspeha":

$$P(N=n \mid x,p) = \binom{n-1}{x-1} p^x (1-p)^{n-x} \text{ za } 0 < x \leq n$$

Očekivana vrednost je $E(N) = x/p$ a varijansa $\sigma^2 = [x(1-p)]/ p^2$

Geometrijska distribucija

Daje verovatnoću postizanja jednog (tj. prvog) "uspeha" u n pokušaja:

$$P(N=n \mid p) = p\ (1-p)^{n-1}$$

Očekivana vrednost je $E(N) = 1/p$ a varijansa $\sigma^2 = (1-p)/ [(1-p)-1]$

Hipergeometrijska raspodela

Kada je data konačna populacija od N elemenata koji su podeljeni u k međusobno isključivih i iscrpnih klasa verovatnoća da izvučemo slučajni uzorak bez vraćanja veličine n koji sadrži n_1 slučajeva jedne klase, n_2 druge i n_k klase k je:

$$P(n=n_1+n_2+\ldots+n_k) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_k}{n_k}}{\binom{N}{n}}$$, pri čemu je $n_1+n_2+\ldots+n_k = n$ dok je $N_1+N_2+\ldots+N_k=N$

Multinomijska distribucija

Ako postoji k međusobno isključivih i iscrpnih događaja sa verovatnoćama $p_1, p_2,\ldots,p_k$ verovatnoća da tačno bude n_1 događaja jedan n_2 događaja dva... n_k događaja k-tih u n nezavisnih i nasumi;nih opservacija:

$$\frac{n!}{n_1!n_2!..n_k!}(p_1)^{n1}(p_2)^{n2}\ldots(p_k)^{nk}$$

gde je $n_1+n_2+\ldots+n_k = n$

Zapamtite: Kada je $n_1 = x$; $n_2 = n - x$, a $k = 2$ multinomijska distribucija se svodi na binomnu.

Puasonova raspodela:

$$P_{(x)} = (e^{-\lambda} \lambda^{x}) / x!$$

Parametar ove distribucije je λ, aritmetička sredina $\mu = \lambda$ a $\sigma^2 = \lambda$. Koeficijent asimetrije je $1/\lambda$ a koeficijent izduženosti $1/\lambda$. Kada $\lambda \to \infty$ koeficijenti asimetrije i izduženosti teže 0.

Napomena: Binomna raspodela teži Puasonovoj kada $p \to 0$ i $n \to \infty$, a np ostaje konačno:

$$\lim_{n\to\infty}\binom{n}{x} p^x q^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Uniformna (rektagularna) funkcija gustine

$$f(x) = \begin{cases} 0 \mid X \le x_1 \\ \dfrac{1}{x_2 - x_1} \mid x_1 < X \le x_2 \\ 0 \mid X > x_2 \end{cases}$$

Aritmetička sredina rektangualrne funkcije gustine je $\mu = (x_1 + x_2) / 2$, a varijansa $\sigma^2 = (x_2 - x_1)^2 / 12$.

Normalna ili Gaus - Bernulijeva (Gauss, Bernoulli) funkcija gustine

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

gde jeX- μ (odstupanje vrednosti slučajne varijable od njene aritmetičke sredine), σ standardna devijacija slučajne varijable, p=3,1415(Ludolfov broj) a e = 2,7183 (osnova prirodnog, Neperovog logaritma).

Teorijski razlog nastanka normalne funkcije gustine:
Ako na neku pojavu deluje vrlo veliki broj nezavisnih činilaca onda je raspored nužno definisan funkcijom normalne distribucije
Empirijski argumenti: Kada se posmatra dovoljno veliki broj pojava u prirodi i društvu uvek postoji tendencija da se rezultati grupišu oko prvog momenta simetrično na obe strane sa pregibima (kada tangenta prelazi u sekantu) na obema stranama koji odgovaraju korenu iz drugog momenta tj sigmi.

Parametri ove funkcije su μ (aritmetička sredina slučajne varijable X) i σ^2(varijansa varijable).

Najvažnije karakteristike normalne raspodele (normalne krive):

1. Zvonasta, unimodalna i simetrična u odnosu na najvišu ordinatu;
2. Aritmetička sredina,Medijana i Mod su jednaki;
3. Kriva se asimptotski približava X-osi;
4. Ukupna površina ispod krive je kao i kod svake krive gustine verovatnoće jednaka 1.
5. 50% površine je levo a 50% desno od najviše ordinate;
6. Površina ispod krive je verovatnoća da slučajna varijabla uzme neku vrednost u intervalu;
7. Normalna distribucija je definisana parametrima μ i σ pa postoji čitava familija normalnih rasporeda zavisno od mogućih vrednosti μ i σ;
8. Između ordinata $\mu \pm \sigma$ nalazi se 68.26 ukupne površine, u intervalu od $\mu \pm 2\sigma$ je 95.44% površine, a interval $\mu \pm 3\sigma$ obuhvata 99.74% ukupne površine.

Normalna distribucija je teorijski model koji, po pretpostavci, dobro opisuje mnoge pojave kojima se psihologija bavi (mada se empirijske ditribucije frekvencija ne poklapaju sa njom sasvim). Nekada se mislilo da za sve prirodne i socijalne fenomene važi zakon normalne raspodele (otuda joj i naziv!). Danas se zna da to nije uvek tako. Normalna distribucija je osnovni model parametrijske statistike zaključivanja. Normalna funkcija gustine se uzima kao pogodan model (aproksimacija) empirijskih distribucija ali empirijske distribucije dobijene na uzorku manje ili više odstupaju od nje.

Standardizovana normalna funkcija gustine

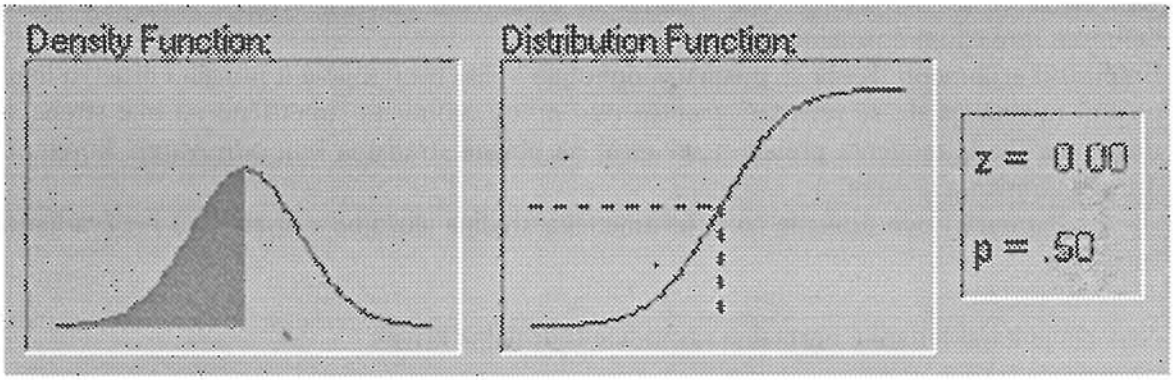
Budući da postoji čitava familija normalnih distribucija (zavisno od μ i σ;) a da je za računanje verovatnoće da normalna slučajna varijabla uzme neku vrednost u određenom intervalu potrebno izračunati određeni integral a to nije baš lako izvesti. Stoga je potrebno formulisati model normalne ditribucije na koji se sve normalne distribucije mogu svesti. To se postiže pravljenjem standardizovane normalne ditribucije za koju postoje gotove tablice

verovatnoća. Standardizovana normalna distribucija je distribucija z-skorova tako da je njena aritmetička sredina 0 a standardna devijacija 1.

Ako μ i σ zamenimo sa 0 i 1 obrazac za normalnu raspodelu postaje:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

Grafički prikaz standardizovane normalne funkcije gustine (Density Function) i njene funkcije distribucije (Distribution Function) izgleda ovako:



(Grafik je preuzet sa URL adrese: http://www.statsoft.com. Odlaskom na ovu URL adresu moguće je posmatrati animaciju koja prikazuje kako se menja površina na funkciji gustine i tačka na ordinati na funkciji distribucije sa promenom z vrednosti.)

Standardni skor

Dobija se po obrascu:

$$z = (X - \mu)/\sigma$$

U psihologiji se z-skorovi koriste, između ostalog za određivanje položaja pojedinca u grupi, kvantitativno klasifikovanje, kategorisanje i selekcionisanje ispitanika i u mnogim metodama statističkog zaključivanja.

Zapamtite: Ako X ima normalnu funkciju gustine sa aritm. sredinom μ i varijansom σ^2, onda z ima takođe normalnu funkciju gustine sa $\mu_z = 0$ i $\sigma_z = 1$.

****

Za određivanje verovatnoće da slučajna varijabla, koja ima normalnu funkciju gustinu sa parametrima μ i σ, uzme neku vrednost u intervalu od $-\infty$ do neke određene vrednosti (označimo je sa x_i) u SPSS programu treba odabrati meni **Transform/Compute**. U polje **Target Variable** treba uneti ime varijable u koju će biti upisane ove verovatnoće. U polje **Numeric Expression** treba uneti iz spiska funkcija funkciju **CDF.NORMAL(q,mean, stddev)** .

Umesto znakova pitanja u ovoj funkciji na mesto argumenta **q** unosi se ili ime varijable sa

rezultatima ili neka određena vrednost. Umesto **mean** i **stddev** treba uneti vrednosti parametara za normalnu funkciju gustine (određene vrednosti za μ i σ).
U istom meniju korišćenjem funkcije **CDFNORM (zvalue)** moguće je odrediti verovatnoću da slučajna varijabla koja ima standardizovanu normalnu funkciju gustinu sa parametrima 0 i 1 uzme neku vrednost u intervalu od $-\infty$ do z_i. Umesto **zvalue** u ovoj funkciji treba upisati određenu z vrednost ili ime neke standardizovane varijable iz fajla sa podacima. Ukoliko se, na primer umesto **zvalue** unese vrednost 1.96 onda će rezultat ove funkcije biti 0.975.
Funkcijom **IDF.NORMAL (p, mean, stddev)** može se odrediti vrednost za koju važi da je verovatnoća da varijabla sa normalnom funkcijom gustine uzme neku vrednost jednaku ili manju datoj vrednosti (tj. od $-\infty$ do date vrednosti). Umesto **p** se upisuje željena verovatnoća a umesto **mean** i **stddev** parametri normalne funkcije gustine (određene vrednosti za μ i σ).
Ukoliko je reč o standardizovanoj funkciji gustine sa parametrima 0 i 1 umesto **IDF.NORMAL** se može koristiti funkcija **PROBIT (prob)** u kojoj se kao **prob** unosi željena verovatnoća.
Ukoliko je potrebno odrediti verovatnoću da slučajna varijabla sa normalnom funkcijom gustine uzme neku vrednost u intervalu od $x_1(z_1)$ do x_2 (z_2) potrebno je naći vrednosti fumkcija za $x_1(z_1)$ i $x_2(z_2)$ pa dobijene vrednosti oduzeti.
Na primer ako vrednost funkcije **CDF.NORMAL(19,15.5,5.4)** iznosi 0.74 a vrednost funkcije **CDF.NORMAL (14,15.5,5.4)** bude jednaka 0.39 verovatnoća da slučajna normalno distribuirana varijabla sa parametrima μ =15.5 i σ =5.4 uzme neku vrednost u intervalu od 14 do 19 je 0.74-0.39 = 0.35.
CDF.*distribution* i **IDF.*distribution*** funkcije postoje i za mnoge druge funkcije gustine. Umesto **NORMAL** u takvim funkcijama stoji ime druge funkcije gustine (npr. **CDF.CHISQ(q,df)** i **IDF.CHISQ(p,df)** za Hi-kvadrat funkciju gustine, **CDF.T(q,df)** i **IDF.T(p,df)** za Studentovu t-funkciju gustine. Princip rada sa ovim funkcijama isti je kao sa funkcijama za normalnu distribuciju s tom razlikom što su parametri ovih funkcija normalno drugačiji. Tako za Hi-kvadrat i t-funkciju treba umesto **df** upisati stepene slobode koji predstavljaju parametar datih funkcija[1].

* *

Hi-kvadrat funkcija gustine

$$f(\chi^2)=\frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(\chi^2\right)^{\frac{\nu-2}{2}} e^{\frac{-\chi^2}{2}}$$

Parametar ove funkcije gustine je ν, tj. stepeni slobode. Očekivanje Hi-kvadrat distribucije je $E(\chi^2_{(\nu)}) = \nu$, a varijansa $\sigma^2_{(\chi 2(\nu))} = 2\nu$. Koeficijent asimetrije je $8 / \nu$, a koeficijent izduženosti $12 / \nu$. Očigledno, sa povećanjem stepeni slobode koeficijenti asimetrije i izduženosti teže nuli, tj. hi - kvadrat distribucija postaje simetričnija, a po izduženosti bliža normalnoj.

[1] Podrobnije o ovome može se pročitati u knjizi N. Vukovića PC verovatnoća i statistika.

Zapamtite:

1. Ako slučajna varijabla $\chi^2_{(\nu 1)}$ ima Hi-kvadrat funkciju gustine sa ν_1 stepeni slobode a nezavisna slučajna varijabla $\chi^2_{(\nu 2)}$ ima Hi-kvadrat funkciju gustine sa ν_2 stepeni slobode onda nova slučajna varijabla dobijena njihovim sabiranjem $\chi^2_{(\nu 1+\nu 2)} = \chi^2_{(\nu 1)} + \chi^2_{(\nu 2)}$ ima Hi-kvadrat funkciju gustine sa $\nu_1+\nu_2$ stepeni slobode.
2. Ako X slučajna varijabla ima normalnu funkciju gustine i aritmetičku sredinu μ i varijansu σ^2 onda slučajna varijabla $Z^2 = [(X- \mu) /\sigma]^2$ ima hi-kvadrat funkciju gustine sa 1 stepenom slobode.
3. Ako je $(x_1,x_2,..,x_k,...,x_n)$ slučajni uzorak iz populacije koja ima normalnu funkciju gustine sa aritmetičkom sredinom μ i varijansom σ^2 i ako $Z_k=(X_k-\mu)/\sigma$ za svaku vrednost k, slučajna varijabla $\chi^2 = \sum_{k=1}^{n} Z_k^{\,2} = \sum_{k=1}^{n} \frac{(X_k-\mu)^2}{\sigma^2}$ ima hi-kvadrat funkciju gustine sa n stepeni slobode, tj sa očekivanom vrednošću n i varijansom 2n.
4. Kada je $E(\chi^2)$, tj. broj stepeni slobode velik $\sqrt{2\chi^2}$ ima približno normalnu funkciju gustine sa aritmetičkom sredinom $\sqrt{(2\,E(\chi^2) -1)}$ i varijansom 1. Prema tome $\sqrt{2\chi^2} - \sqrt{(2\,E(\chi^2) -1)}$ ima približno normalnu funkciju gustine sa $\mu = 0$ i $\sigma^2 = 1$ za veliki broj stepeni slobode.

Studentova t funkcija gustine

Ova distribucija se najčešće primenjuje u testiranju značajnosti razlika između aritmetičkih sredina, tj. pri korišćenju t - testa.
Ako slučajna varijabla X ima normalnu funkciju gustine sa aritm. sredinom μ i varijansom σ^2 ali varijansa populacije nije poznata već se ocenjuje na osnovu varijanse dobijene na uzorku:

$$S^2 = \sum_{i=1}^{n} (x_i - M)^2 * (n-1)^{-1}$$

onda slučajna varijabla t, takva da je $t = (M - \mu)/ SE_M$ (gde je SE_M ocena standardne greške za aritmetičku sredinu dobijena na osnovu $S^2/\sqrt{n}$ a ne $\sigma^2 / \sqrt{n}$) ima funkciju gustine:

$$f(t) = G(\nu)\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2}$$

pri čemu je $G(\nu) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$ konstanta za određeno ν, a ν je parametar, tj. broj

stepeni slobode.[13] Dakle, t funkcija gustine je u stvari, funkcija gustine slučajne varijable koja je količnik dveju nezavisnih slučajnih varijabli: slučajne varijable koja ima standardizovanu normalnu funkciju gustine i kvadratnog korena količnika varijable koja ima hi - kvadrat funkciju gustine sa ν stepeni slobode i odgovarajućeg broja stepeni slobode, tj. ν.
Jednostavnije rečeno, t varijabla :

$$t = \frac{Z}{\sqrt{\frac{H_\nu}{\nu}}},$$

(pri čemu je $Z \sim N(0, 1)$, $H_\nu \sim \chi^2_{(\nu)}$, a ν broj stepeni slobode) ima Studentovu t distribuciju sa parametrom ν, tj sa ν stepeni slobode. (Oznaka ~ je u ovom slučaju skraćenica za "ima funkciju gustine... ").

Za $\nu > 1$ je $E(t) = 0$, a za $\nu > 2$ je $\sigma^2(t) = \nu / (\nu-2)$. Dakle, sa porastom ν se σ^2_t približava 1 jer je

$$\lim_{\nu \to \infty} \frac{\nu}{\nu - 2} = 1$$

Koeficijent asimetrije t distribucije je 0, a koeficijent izduženosti je 6 / (n-4). Kada n teži ∞ koeficijent izduženosti teži 0. Na osnovu svega toga, očigledno je da sa porastom ν Studentova distribucija teži normalnoj.

Zapamtite: Sa povećanjem broja stepeni slobode Studentova distribucija teži normalnoj.

Snidikorova f funkcija gustine

Ova se distribucija najviše koristi u postupcima analize varijanse. Slučajna varijabla F koja ima Snidikorovu funkciju gustine je količnik dveju nezavisnih varijabli H_1 i H_2, pri čemu svaka od tih varijabli ima hi - kvadrat dustribuciju i podeljena je odgovarajućim brojem stepeni slobode. Tačnije i jednostavnije, slučajna varijabla F:

$$F = \frac{\frac{H_1}{\nu_1}}{\frac{H_2}{\nu_2}},$$

[13] $\Gamma(p)$je gama funkcija (Ojlerova funkcija druge vrste) koja je definisana na sledeći način:

$$\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt, \quad p > 0$$

Ova funkcija ma sledeće osobine:

1. $\Gamma(p+1) = p\,\Gamma(p)$. Ako je p prirodna broj p = n pa je $\Gamma(n+1) = n!$
2. $\Gamma(1/2)=\sqrt{\pi}$: $\Gamma(n+1/2)=[(2n-1)/2]*[(2n-3)/2]*\ldots \sqrt{\pi}$
3. $\Gamma(p)*\Gamma(1-p)= \pi/\sin p\pi$, $0<p<1$
4. Za $n < 0$: $\Gamma(n) = \Gamma(n+1)/n$.

(pri čemu su H_1 i H_2 nezavisne, $H1 \sim \chi^2_{(\nu 1)}$, $H_2 \sim \chi^2_{(\nu 2)}$, a ν_1 i ν_2 su stepeni slobode za H_1 i H_2),
ima Snidikorovu funkciju gustine sa parametrima ν_1 i ν_2:

$$f(F) = \frac{\nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}}}{B(\nu_1, \nu_2)} F^{\frac{\nu_1}{2}-1} (\nu_2 + \nu_1 F)^{\frac{-(\nu_1+\nu_2)}{2}}$$

$E(F) = \nu_2 / \nu_2 - 2$, a varijansa $\sigma^2(F) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$

Zapamtite: Ako slučajna varijabla ima T funkciju gustine sa ν stepeni slobode onda t^2 ima F funkciju gustine sa $\nu_1=1$ i $\nu_2 = \nu$ stepeni slobode.

III POVEZANOST TEORIJSKIH DISTRIBUCIJA VEROVATNOĆA I EMPIRIJSKIH ISTRAŽIVANJA POJAVA

Osnovni cilj statističkog zaključivanja (pa i većine istraživanja) je zaključivanje o populaciji na osnovu rezultata dobijenih na reprezentativnom uzorku izvučenom iz date populacije. Da bismo vršili statističko zaključivanje o populaciji na osnovu ispitivanja uzorka nužno koristimo određene matematičke modele (teorijske distribucije ili zakone verovatnoće) koji po pretpostavci dobro opisuju empirijsku distribuciju rezultata dobijenih u istraživanju.Budući da po pretpostavci dati matematički modeli dobro reprezentuju pojedine fenomene koje ispitujemo sledeći logiku matematičkog modela donosimo zaključke o populaciji u pogledu ispitivane pojave.

Tri vrste distribucija koje su važne za razumevanje statističkog zaključivanja su:

1. Distribucija populacije
2. Distribucija uzorkovanja i
3. Distribucija uzorka.

Distribucija populacije ima efekat na preostale dve ali je kritično poznavanje ditribucije uzorkovanja jer bez njenog poznavanja zaključivanje nije moguće.

Empiriska distribucija učestalosti

Odnosi se na uzorak a dobija se prostim tabuliranjem vrednosti koje slučajna varijabla uzima na uzorku. Ako je uzorak slučajan, relativna frekvenca za svaku vrednost slučajne varijable dobijena na uzorku je ocena verovatnoće ove vrednosti u populaciji.
Distribucija frekvenci determinisana je :

1. prirodom distribucije verovatnoća ili funkcije gustine varijable u populaciji i
2. varijabilnošću procesa slučajnog uzorkovanja

III. 1. Odnos matematičkog modela i prirode pojave na koju se on primenjuje

Između matematičkih pojmova ili modela i prirodnih pojava postoji izomorfizam (sličnost oblika). Primenjujući neki matematički model na pojave (usled dovoljnog slaganja oblika pojava i oblika matematičkih izraza) možemo doći do vrlo korisnih pretpostavki o prirodi pojave sledeći logiku njenog matematičkog modela. Naravno, pretpostavke do kojih dolazimo na ovaj način potrebno je podvrgnuti empirijskoj proveri.

III. 2. Odnos teorijskih distribucija verovatnoće i empirijskih distribucija učestalosti

Teorijske distribucije verovatnoća i funkcije gustine verovatnoće vezane su za populaciju. Distribucija verovatnoća za diskretnu slučajnu varijablu pokazuje verovatnoću da slučajno izabrani član populacije ima određenu vrednost u pogledu varijable koja se meri.
Za kontinuiranu slučajnu varijablu funkcija gustine verovatnoće dopušta da se odredi verovatnoća da član populacije slučajno izabran ima vrednost koja će se nalaziti između određenih granica.
Kada se radi o uzorku tabuliranje podataka da bi se pokazala učestalost sa kojima se svaka vrednost varijable pojavljuje u uzorku daje funkciju frekvencije ili distribuciju frekvencija. Distribucija frekvencija u uzorku određena je prirodom distribuije verovatnoća date varijable u populaciji i varijabilnošću procesa izbora reprezentativnog uzorka.Distribucija frekvencija će, prema tome, odstupati slučajno manje ili više od distribucije populacije.

****

Korisnici Priručnika koji žele da prodube svoje razumevanje osnovnih pojmova teorije verovatnoće i teorijske statistike koji su veoma sažeto prikazani u ovom dodatku mogu da konsultuju neku od knjiga koje su u spisku literature pod brojevima 6, 10, 17, 18 i 22.

DODATAK 3: OSNOVNI POJMOVI MATRIČNE ALGEBRE

Osnovna namena Dodatka 3 nije da korisnika Priručnika obavesti na matematički sasvim korektan način o matričnoj algebri, već da mu omogući da prati matričnu notaciju koja je često korišćena u Priručniku. Autor je nastojao da osnovne pojmove iz matrične algebre prikaže što jednostavnije i da, pri tom, što je više moguće, koristi primere koji su u bliskoj vezi sa sadržajem Priručnika. Korisnici Priručnika koji žele da postignu dublje razumevanje matrične algebre trebalo bi da, pored ovog Dodatka, prouče dodatak Elementi linearne algebre u knjizi K. Momirovića, B. Wolfa i D. A. Popovića *Uvod u teoriju merenja I.* Vrlo korisne knjige na engleskom jeziku za savladavanje elemenata matrične i linearne algebre koje su prilagođene čitaocu koji nema primarno matematičko obrazovanje su:

- Green, P.E & Carroll, J.D. (1976). *Mathematical Tools for Applied Multivariate Analysis*, London: Academic Press;

- Horst, P. (1963). *Matrix Algebra for Social Scientists*, New York: Holt.

- Searle, S. R. (1966). *Matrix Algebra for The Biological Sciences*, New York: John Wiley & Sons, Inc.

Na prezentaciji predmeta Statistika u psihologiji mogu se pronaći linkovi za prezentacije iz ove oblasti
(URL adresa: **http://www.f.bg.ac.yu/psihologija/statistika/links.htm**) koje sadrže poglavlja ili čitave knjige iz linearne i matrične algebre koje su napisane veoma razumljivo.

SADRŽAJ DODATKA 3

1. MATRICE 3

1.1. Kvadratna matrica 4

1.2. Dijagonalna matrica 4

1.3. Skalarna matrica 4

1.4. Matrica identiteta (I) 4

1.5. Množenje matrice skalarom 4

1.6. Transpon matrice (X^t) 4

2. VEKTORI 5

2. 1. Sabiranje i oduzimanje matrica i vektora 6

2.2. Skalarni proizvod vektora 6

2.3. Dužina vektora 7

2.4. Jedinični vektor 7

2.5. Vektor jedinica ili sumacioni vektor 7

2.6. Linearna kombinacija vektora 7

2.7. Ortogonalnost dva vektora 7

2.8. Linearna nezavisnost vektora 7

3. OBIČNI PROIZVOD MATRICA 8

4. SIMETRIČNE MATRICE 11

5. TRAG MATRICE 11

6. DETERMINANTA MATRICE 12

7. REGULARNI INVERZ MATRICE 12

8. ORTOGONALNE MATRICE 13

9. RANG MATRICE 14

10. SVOJSTVENE VREDNOSTI I SVOJSTVENI VEKTORI MATRICE 14

11. KVADRATNA FORMA 15

1. MATRICE

Matrica je matematička struktura pravougaonog ili kvadratnog oblika koja sadrži nizove elemenata poređanih u redove i kolone. Broj redova i broj kolona koji matrica sadrži govore o redu ili dimenzijama matrice. Tako, za matricu koja ima 5 redova i 6 kolona kažemo da je matrica reda 5 x 6 (pet puta šest). Matrice se obično obeležavaju velikim zadebljanim slovima kao, na primer, **X**, **A**, **D**, Σ, Δ...

Svaki element matrice nalazi se na mestu na kojem se ukršta određeni red i određena kolona matrice. U ovom tekstu pretpostavljaćemo da su elementi koji se nalaze u matrici realni brojevi, mada u opštem slučaju to mogu biti elementi nekog drugog skupa, na primer skupa kompleksnih brojeva. U opštem slučaju element matrice označava se malim slovom uz koje idu dva indeksa: prvi indeks je oznaka reda, a drugi indeks je oznaka kolone u kojoj se u matrici nalazi taj element. Tako, na primer, u matrici podataka, u kojoj su uobičajeno ispitanici ili jedinice posmatranja u redovima, a varijable u kolonama, rezultat ispitanika L. T. (treći red) na varijabli introverzija (druga kolona) nalazi se na mestu na kojem se ukršta treći red i druga kolona i označava se oznakom x_{32}. Ako matrica podataka izgleda ovako:

$$\mathbf{X} = \begin{bmatrix} 15 & 12 & 10 \\ 18 & 11 & 9 \\ 12 & 4 & 11 \\ 16 & 2 & 18 \\ 14 & 8 & 12 \end{bmatrix}$$

onda je reč o matrici reda 5 x 3, a rezultat ispitanika L. T. na varijabli introverzija jednak je 4. To bismo mogli još napisati i ovako: $x_{32} = 4$. Elementi koje sadrži matrica uobičajeno se pišu unutar uglaste zagrade kao što je to prikazano u ovom primeru.

U opštem slučaju, matrica reda n x m izgledala bi ovako:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \cdots x_{1j} \cdots x_{1m} \\ x_{21} & x_{22} & x_{23} \cdots x_{2j} \cdots x_{2m} \\ x_{31} & x_{32} & x_{33} \cdots x_{3j} \cdots x_{3m} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{i1} & x_{i2} & x_{i3} \cdots x_{ij} \cdots x_{im} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n1} & x_{n2} & x_{n3} \cdots x_{nj} \cdots x_{nm} \end{bmatrix}$$

Uočimo da opšti element x_{ij} ima dva indeksa (indeks reda i indeks kolone u kojoj je element) i da indeks i (indeks reda) uzima vrednosti od 1 do n, a da se vrednosti indeksa j (indeksa kolone) kreću od 1 do m. Moguće je, stoga, matricu **X** napisati i u sledećem obliku:

$$\mathbf{X} = \{x_{ij}\}, \quad i = 1, 2,...,n; \quad j = 1, 2,...,m.$$

1.1. Kvadratna matrica

Matrica koja ima jednak broj redova i kolona zove se kvadratnom matricom. Tipični primeri kvadratnih matrica u analizi podataka u psihologiji su matrica kovarijansi i matrica interkorelacija.[1] Na primer, matrica interkorelacija za četiri varijable (MSUKUP, E, NEUROT, EKSTRA) izgledala bi ovako:

	MSUKUP	E	NEUROT	EKSTRA
MSUKUP	1.00	.32	.53	-.32
E	.32	1.00	-.05	.06
NEUROT	.53	-.05	1.00	-.41
EKSTRA	-.32	.06	-.41	1.00

Ova matrica je reda 4 x 4.

1.2. Dijagonalna matrica

Matrica čiji su svi vandijagonalni elementi, tj. elementi za koje je $i \neq j$ (na primer x_{12}, x_{13}, x_{14}, x_{21}...) jednaki nuli, a čiji svi dijagonalni elementi, tj. elementi za koje je $i = j$ (na primer x_{11}, x_{22}, x_{33}, x_{44}...) ne moraju biti jednaki nuli je dijagonalna matrica.

1.3. Skalarna matrica

Skalarna matrica je dijagonalna matrica čiji su svi dijagonalni elementi jednaki.

1.4. Matrica identiteta (I)

Skalarna matrica čiji su svi dijagonalni elementi jednaki jedinici zove se matricom identiteta, uobičajeno u oznaci I. Kada bi svi elementi izvan glavne dijagonale u matrici interkorelacija u prethodnom primeru bili jednaki nuli onda bismo u tom slučaju imali primer matrice identiteta.

1.5. Množenje matrice skalarom

U matričnoj algebri, realni brojevi se obično zovu skalarima. Matrica se množi skalarom ili realnim brojem tako što se svaki element matrice pomnoži tim skalarom.

1.6. Transpon matrice ($\mathbf{X}^t$)

Kada kolone neke matrice pretvorimo u redove (pa, prema tome, redove u kolone) onda smo početnu matricu transponovali, a matrica koju dobijamo ovom operacijom je transpon početne matrice. Transpon matrice obično se označava malim slovom t kao eksponentom iznad oznake matrice (što je slučaj u ovom Priručniku) ili jednostrukim apostrofom uz oznaku matrice ($\mathbf{X}'$).

---

[1] Pojmovi kovarijanse i korelacije objašnjeni su u poglavlju XI osnovnog teksta.

Na primer, transpon matrice **X**:

$$\mathbf{X} = \begin{bmatrix} 15 & 12 & 10 \\ 18 & 11 & 9 \\ 12 & 4 & 11 \\ 16 & 2 & 18 \\ 14 & 8 & 12 \end{bmatrix}$$

jednak je:

$$\mathbf{X}^t = \begin{bmatrix} 15 & 18 & 12 & 16 & 14 \\ 12 & 11 & 4 & 2 & 8 \\ 10 & 9 & 11 & 18 & 12 \end{bmatrix}$$

Formalno, operaciju transponovanja mogli bismo definisati na sledeći način:

Ako je $\mathbf{X} = (x_{ij})$, tada je $\mathbf{X}^t = (x_{ij})^t = (x_{ji})$.

Očigledno, element koji je u početnoj matrici bio u i-tom redu i j-toj koloni u transponovanoj matrici biće u j-tom redu i i-toj koloni.

Zapamtite:

- Ako je a skalar, tj. realan broj, tada je $(a\mathbf{X})^t = a\,\mathbf{X}^t$;
- $(\mathbf{X}^t)^t = \mathbf{X}$. Transpon transponovane matrice jednak je početnoj matrici.

2. VEKTORI

Mada se pojam vektora može formalno mnogo bolje definisati, u ovom tekstu pod vektorom ćemo podrazumevati prosto matricu koja ima samo jedan red (redni vektor) ili samo jednu kolonu (kolonski vektor). Dakle, **redni vektor je matrica reda 1 x n, a kolonski vektor je matrica reda m x 1**. Svaka varijabla u matrici podataka na kojoj su podaci za n ispitanika ili jedinica posmatranja u ovom smislu predstavlja kolonski vektor. Isto tako, svaki ispitanik koji ima podatke na m varijabli u toj matrici može se posmatrati kao redni vektor. Kolonske vektore ćemo u ovom tekstu označavati malim zadebljanim slovima (na primer **x**), a redne vektore kao transpone kolonskih vektora $(\mathbf{x}^t)$[2]. Prema tome, budući da su vektori, onako kako su ovde definisani, specijalni slučajevi matrica, za njih važe pravila množenja skalarom i transponovanja koja smo objasnili u prethodnom tekstu.

[2] Korisnik ovog Priručnika treba da ima na umu, ako koristi matematičke knjige iz ove oblasti, da neki autori koriste upravo suprotno označavanje od onoga koje je korišćeno u ovom tekstu.

2.1. Sabiranje i oduzimanje matrica i vektora

Matrice i vektori se mogu sabirati i oduzimati samo ako su istog reda, tj. jednakih dimenzija. U tom slučaju, oduzimanje i sabiranje protiče tako da se saberu (oduzmu) elementi dva vektora ili dve matrice koji imaju isti indeks. Figurativno rečeno, treba u mislima "poklopiti" dva vektora ili dve matrice istih dimenzija i sabrati (oduzeti) elemente koji poklapaju jedan drugi.

Zapamtite: Sabiranje (oduzimanje) vektora ili matrica, kao i množenje matrice ili vektora skalarom, su **komutativne** operacije jer važe sledeća pravila:

$$\mathbf{X} + \mathbf{Y} = \mathbf{Y} + \mathbf{X}; \qquad a\mathbf{x} = \mathbf{x}a; \qquad a\mathbf{X} = \mathbf{X}a$$

2.2. Skalarni proizvod vektora

Da bismo dobili skalarni proizvod dva vektora prvi član proizvoda mora biti redni vektor, a drugi član proizvoda mora biti kolonski vektor. Pri tome, broj elemenata u rednom vektoru mora biti jednak broju elemenata u kolonskom vektoru. Množenje se odvija tako što se, figurativno rečeno, "poklope" elementi dva vektora koji imaju isti redosled (prvi član jednog vektora sa prvim članom drugog vektora i tako redom), zatim se elementi dva vektora koji se "poklapaju" pomnože i, na kraju, tako dobijeni proizvodi saberu. Kao rezultat ovoga dobija se jedan broj, skalar. Otuda i potiče naziv ovog proizvoda. Ponekada se ovaj proizvod zove i unutrašnjim proizvodom vektora.

Najjednostavnije se skalarni proizvod dva vektora ($\mathbf{x}^t$ i $\mathbf{y}$) može definisati na sledeći način:[3]

Ako su $\mathbf{x}^t = (x_1, x_2, \ldots, x_n)$ i $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ vektori reda 1 x n i n x 1

tada je $\mathbf{x}^t\,\mathbf{y} = (x_1 * y_1 + x_2 * y_2 + x_3 * y_3 + \cdots + x_n * y_n) = \sum_{i=1}^{n} (x_i * y_i)$

skalarni proizvod vektora **x** i **y**.

Primer:

Ako su $\mathbf{x} = \begin{bmatrix} 2 \\ 5 \\ 1 \\ 3 \end{bmatrix}$ i $\mathbf{y} = \begin{bmatrix} 5 \\ 4 \\ 6 \\ 4 \end{bmatrix}$ dva vektora, njihov skalarni proizvod je $\mathbf{x}^t\,\mathbf{y}$ = (2*5 + 5*4 + 1*6 + 3*4) = 48. Uočimo da je potrebno vektor **x** transponovati pre skalarnog množenja.

---

[3] Operator sabiranja Σ koji se pojavljuje u ovoj definiciji objašnjen je u Dodatku 4 (strana 6).

2.3. Dužina vektora

Dužina vektora jednaka je kvadratnom korenu iz zbira kvadriranih elemenata vektora. Uočimo da se dužina vektora može definisati i preko skalarnog proizvoda vektora sa samim sobom. Naime, ako je $\mathbf{x}$ vektor reda 1 x n, tada je

$$\mathbf{x}^t\mathbf{x} = (x_1*x_1 + x_2*x_2 + x_3*x_3 + \cdots + x_n*x_n) = \sum_{i=1}^{n}(x_i^2),$$

pa je, prema tome, dužina vektora jednaka kvadratnom korenu iz skalarnog proizvoda vektora sa samim sobom.

2.4. Jedinični vektor

Vektor čija je dužina jednaka jedinici ili vektor kod kojeg je zbir kvadriranih elemenata jednak jedinici zove se jedinični vektor. (Ove dve definicije su iste jer je kvadratni koren iz broja 1 jednak jedinici).

2.5. Vektor jedinica ili sumacioni vektor

Sumacioni vektor je vektor čiji su svi elementi jednaki jedinici. U Priručniku je ovaj vektor uobičajeno označen slovom $\mathbf{e}$.

Zapamtite: Skalarni proizvod sumacionog vektora sa nekim drugim vektorom $\mathbf{x}$ daje rezultat koji je jednak zbiru elemenata vektora $\mathbf{x}$:

$$\mathbf{e}^t\mathbf{x} = \sum_{i=1}^{n} x_i .$$

2.6. Linearna kombinacija vektora

Vektori se mogu linearno kombinovati da bi se dobio novi vektor. To se radi tako što se svaki vektor koji ulazi u linearnu kombinaciju pomnoži nekim brojem, tj. skalarom, a zatim se dobijeni proizvodi saberu.
Jednostavnije rečeno, ako su $a_1, a_2, \ldots, a_m$ realni brojevi, a $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ skup m vektora, onda se linearna kombinacija ovih vektora može definisati na sledeći način:

$$\sum_{j=1}^{m} a_j * \mathbf{x}_j = a_1 * \mathbf{x}_1 + a_2 * \mathbf{x}_2 + \cdots + a_m * \mathbf{x}_m = \mathbf{y}$$

Dakle, vektor $\mathbf{y}$ nastao je linearnim kombinovanjem skupa vektora $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, a brojevi $a_1, a_2, \ldots, a_m$ zovu se i koeficijentima za linearnu kombinaciju.

2.7. Ortogonalnost dva vektora

Dva vektora $\mathbf{x}$ i $\mathbf{y}$ su ortogonalni akko ("ako i samo ako") je njihov skalarni proizvod ($\mathbf{x}^t\mathbf{y}$) jednak nuli.

2.8. Linearna nezavisnost vektora

Ako se linearnim kombinovanjem nekog skupa vektora može dobiti vektor nula, tj. vektor čiji su svi elementi nule ali tako da se bar jedan vektor koji ulazi u

linearnu kombinaciju može pomnožiti brojem različitim od nule, onda je taj skup vektora linearno **zavisan**.

Jednostavnije rečeno, ako su $a_1, a_2, \ldots, a_m$ realni brojevi, a $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ skup m vektora, ovaj skup vektora je linearno nezavisan, ako je moguće samo trivijalno rešenje sledeće jednačine:

$$\sum_{j}^{m} a_j * \mathbf{x}_j = a_1 * \mathbf{x}_1 + a_2 * \mathbf{x}_2 + \cdots + a_m * \mathbf{x}_m = \mathbf{0}$$

Trivijalno rešenje ove jednačine je rešenje pri kojem su svi koeficijenti za linearnu kombinaciju jednaki nuli, tj. $a_1 = a_2 = \cdots = a_m = 0$.

3. OBIČNI PROIZVOD MATRICA

Postoje različiti tipovi množenja dveju matrica. U ovom tekstu objasnićemo samo tzv. obični proizvod dveju matrica koji se može izvesti samo onda kada matrica koja je u proizvodu prva (sa leve strane) ima onoliko kolona koliko redova ima matrica koja je u proizvodu druga (sa desne strane). Ovaj proizvod se zapravo svodi na skalarno množenje redova prve matrice sa kolonama druge matrice.Dakle, ako je matrica **X** reda n x m, a matrica **Y** reda m x p onda se ove matrice mogu pomnožiti.

Tada je:

$$\mathbf{X}_{n \times m}\, \mathbf{Y}_{m \times p} = \mathbf{Z}_{n \times p}$$

Uočimo da indeksi matrica koje se množe, a koji se nalaze sa unutrašnje strane (m) moraju biti isti da bi se matrice mogle pomnožiti. Isto tako, važno je uočiti da "spoljašnji" indeksi dveju matrica (n i p) određuju red ili dimenzije matrice koja se dobija njihovim množenjem. Obični proizvod dveju matrica koje se mogu pomnožiti dobija se tako što se prvi red matrice koja je sa leve strane skalarno množi sa svakom kolonom matrice koja je sa desne strane. Brojevi koji se dobiju na ovaj način smeštaju se redom u prvi red matrice koja se dobija množenjem dveju matrica. Dakle elementi nove matrice, koji se dobijaju skalarnim množenjem određenog reda matrice sa leve strane i određene kolone matrice sa desne strane, imaju prvi indeks koji odgovara indeksu reda leve matrice, dok drugi indeks tog elementa odgovara indeksu kolone desne matrice. Isti postupak se ponavlja sa svakim sledećim redom matrice sleva.

Jednostavnije se ovaj produkt može definisati na sledeći način:

Ako je **X** matrica reda n x m, a matrica **Y** reda m x p, onda je matrica **Z** reda n x p proizvod matrica **X** i **Y**, tj. **Z** = **XY**, pri čemu je element u i-tom redu i j-toj koloni matrice **Z** dat sledećim izrazom:

$$z_{ij} = \sum_{k=1}^{m} (x_{ik} * y_{kj})$$

Primer:

Ako su date matrice **X** i **Y**, pri čemu je:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 2 \\ 3 & 1 & 1 \\ 1 & 2 & 1 \\ -1 & 3 & 2 \end{bmatrix}; \mathbf{Y} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$$

tada je matrica **Z**, pri čemu je:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 3 & 6 \\ 1 & 3 \\ -1 & -1 \end{bmatrix}$$

jednaka proizvodu matrica **X** i **Y**, tj. **Z** = **XY**.

Kako se izvodi ovo množenje?

1. Pomnožimo skalarno prvi red matrice **X** sa prvom kolonom matrice **Y**:
   1*1 + 0*0 + 2*0 = 1. Broj koji smo dobili je element u prvom redu i prvoj koloni matrice **Z**;
2. Pomnožimo skalarno prvi red matrice **X** sa drugom kolonom matrice **Y**:
   1*2 + 0*1 + 2*-1 = 0. Broj koji smo dobili je element u prvom redu i drugoj koloni matrice **Z**;
3. Po istom principu skalarno množimo drugi red matrice **X**, sa prvom, pa sa drugom kolonom matrice **Y**. Dobijeni elementi nalaze se u drugom redu i prvoj, odnosno drugoj koloni matrice **Z**;
4. To isto ponovimo sa trećim i četvrtim redom matrice **X**;

Uočimo, dakle, da elementi za matricu **Z** koje dobijamo na ovaj način imaju prvi indeks koji odgovara indeksu reda leve matrice dok drugi indeks odgovara indeksu kolone desne matrice.

Matrica **X** je reda 4 x 3, a matrica **Y** reda 3 x 2. Matrica koja se dobija njihovim množenjem, matrica **Z** je reda 4 x 2, tj. njen red jednak je broj redova leve x broj kolona desne matrice. To je uvek slučaj u običnom množenju matrica koje smo ovde prikazali.

Veoma je važno uočiti na ovom primeru da se dve matrice mogu pomnožiti u jednom redosledu, ali ne mora biti moguće pomnožiti ih u obrnutom redosledu. Naime, nije moguće u navedenom primeru ni napraviti proizvod **YX** prema pravilu koje važi za obično množenje matrica!

Zapamtite:

- Za množenje matrica ne važi svojstvo komutativnosti!

  U opštem slučaju, $\mathbf{XY} \neq \mathbf{YX}$.

- Množenje matrice **X** matricom identiteta **I** ne menja matricu **X**;

- $(\mathbf{X} + \mathbf{Y})^t = \mathbf{X}^t + \mathbf{Y}^t$. Transpon zbira matrica jednak je zbiru transponovanih matrica;

- $(\mathbf{X} + \mathbf{Y})\,\mathbf{Z} = \mathbf{XZ} + \mathbf{YZ}$; $\mathbf{Z}(\mathbf{X} + \mathbf{Y}) = \mathbf{ZX} + \mathbf{ZY}$. Množenje je operacija distributivna i sleva i zdesna u odnosu na sabiranje matrica

- $(\mathbf{XY})^t = \mathbf{Y}^t\,\mathbf{X}^t$; $(\mathbf{X}_1\,\mathbf{X}_2 \cdots \mathbf{X}_k)^t = \mathbf{X}_k^t \cdots \mathbf{X}_2^t\,\mathbf{X}_1^t$

  Transpon proizvoda dveju ili više matrica jednak je proizvodu transpona ovih matrica u obrnutom redosledu.

Budući da su vektori samo specijalni slučajevi matrica, množenje vektora matricom ili matrice vektorom odvija se po pravilu za množenje matrica.

Primer1:

Ako je $\mathbf{e}^t = (1\ 1\ 1)$ vektor jedinica (sumacioni vektor) reda 1 x 3, a matrica **Y** ista kao u prethodnom primeru, onda je $\mathbf{e}^t\mathbf{Y}$ jednako:

1*1+1*0+1*0 =1;
1*2+1*1+1*-1=2;

$$\text{tj. } \mathbf{e}^t\mathbf{Y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Uočimo da smo ovim množenjem dobili matricu, tj. vektor u kojem su elementi zapravo zbirovi kolona matrice **Y**.

Primer 2:

Ako je **R** matrica interkorelacija tri varijable:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

a vektor $\mathbf{b}^*$ vektor parcijalnih regresionih koeficijenata[4]:

$$\mathbf{b}^* = \begin{bmatrix} b_1^* \\ b_2^* \\ b_3^* \end{bmatrix}$$

tada je proizvod $\mathbf{Rb}^*$ jednak:

$$\mathbf{q} = \begin{bmatrix} 1*b_1^* + r_{12}*b_2^* + r_{13}*b_3^* \\ r_{21}*b_1^* + 1*b_2^* + r_{23}*b_3^* \\ r_{31}*b_1^* + r_{32}*b_2^* + 1*b_3^* \end{bmatrix}$$

Dobili smo, dakle, vektor **q** reda 3 x 1. (U konkretnom slučaju, kao što je poznato iz regresione analize elementi vektora **q** su korelacije prediktorskih varijabli sa kriterijumskom varijablom).

4. SIMETRIČNE MATRICE

U analizi podataka u psihologiji veoma se često sreću simetrične matrice, tj. matrice čiji su odgovarajući elementi sa jedne i druge strane glavne dijagonale jednaki. Naime, element u prvom redu i drugoj koloni takvih matrica jednak je elementu u drugom redu i prvoj koloni i tako redom. Tako izgledaju, na primer, matrice kovarijansi i matrice interkorelacija. Takva je matrica **R** u prethodnom primeru br. 2. Jednostavnije rečeno, simetrične matrice jednake su svom transponu.

Zapamtite:

- Proizvod dveju simetričnih matrica nije nužno simetrična matrica;
- Kao proizvod matrice i njenog transpona u bilo kojem redosledu ($\mathbf{X}^t\mathbf{X}$ ili $\mathbf{X}\mathbf{X}^t$) dobija se simetrična matrica.

5. TRAG MATRICE

Trag kvadratne matrice **X**, u oznaci trag (**X**) ili tr (**X**), jednak je zbiru elemenata na glavnoj dijagonali matrice. Očigledno, trag imaju samo kvadratne matrice.

Ako je matrica **X** reda m x m tada je trag matrice **X**:

$$\mathrm{tr}(\mathbf{X}) = \sum_{i=1}^{m} x_{ii}$$

[4] Na ovaj primer čitalac se može vratiti kada proučava multiplu regresionu analizu (glava XIV)

6. DETERMINANTA MATRICE

Jedna od vrlo važnih skalarnih funkcija matrice je njena determinanta[5]. Determinantu mogu imati samo kvadratne matrice. Budući da je formalna definicija determinante prilično komplikovana, u ovom tekstu ćemo determinantu opisati tako da čitalac stekne samo opštu predstavu o njoj:

Determinanta matrice **X** reda m x m, u oznaci $|\mathbf{X}|$, dobija se tako što se sabere m! (m faktorijel) proizvoda[6], pri čemu svaki proizvod ima pozitivan ili negativan predznak, a svaki proizvod sastavljen je od m elemenata matrice[7]. Isto tako, u jednom istom proizvodu ne smeju biti dva elementa iz istog reda ili iste kolone matrice. Očigledno, vrednost determinante zavisi od svih elemenata u matrici.

Ilustrovaćemo računanje determinante na veoma jednostavnom primeru. Ako je matrica **X**, reda 2 x 2, takva da je:

$$\mathbf{X} = \begin{bmatrix} 6 & 3 \\ 2 & 4 \end{bmatrix}$$

onda je determinanta matrice **X**, $|\mathbf{X}| = 6*4 - 2*3 = 24 - 6 = 18$.
Uočimo da smo sabrali 2! ("dva faktorijel"), tj. 2 proizvoda. Prvi ima pozitivan, a drugi, iz razloga koje nećemo ovde objasniti, negativan predznak. Ni u jednom od ova dva proizvoda nema dva elementa koji su iz istog reda ili iste kolone matrice.
Determinanta matrice je važna jer se, između ostalog, na osnovu vrednosti determinante matrice može ustanoviti da li matrica može imati **regularni inverz** (inverz je objašnjen u nastavku teksta). Ako je determinanta neke matrice jednaka nuli, tj. ako je $|\mathbf{X}| = 0$, matrica **X** je **singularna** i nema regularni inverz.

Kolone matrice podataka možemo posmatrati kao skup varijabli ili vektora. Matrica podataka čije kolone ne predstavljaju linearno nezavisan skup vektora ima determinantu jednaku 0 i ne može imati regularni inverz. Ako, na primer, u matrici podataka, čiji regularni inverz je potrebno naći tokom statističke analize, postoji varijabla koja je zbir nekih od preostalih varijabli u matrici, regularni inverz takve matrice nećemo moći izračunati.

Osim navedenog, determinanta matrice u mnogim multivarijacionim analizama može poslužiti kao mera **generalizovane varijanse** skupa varijabli koje se nalaze u kolonama matrice[8].

7. REGULARNI INVERZ MATRICE

Umesto deljenja kakvo je definisano u skalarnoj algebri, u matričnoj algebri postoji operacija množenja regularnim inverzom matrice. Dakle, umesto da matricu **X** "podelimo" matricom **Y**, u matričnoj algebri zapravo matricu **X** množimo regularnim inverzom matrice **Y**. Regularni inverz matrice označava se obično eksponentom -1

---

[5] Skalarne funkcije matrice preslikavaju elemente matrice u jedan broj, skalar. Dakle, kao rezultat neke funkcije u koju su uključeni elementi matrice dobijamo jedan broj.
[6] m faktorijel, u oznaci m!, jednak je proizvodu celih brojeva od 1 do m. Ako je, na primer, m = 5, tada je 5! = 5*4*3*2*1=120.
[7] Postupak kojim se određuje da li proizvod ima pozitivan ili negativan predznak objašnjen je u udžbenicima linearne algebre.
[8] O ovome se može pročitati u knjizi Z. Kovačića, Multivarijaciona analiza.

nad oznakom matrice. Na primer, regularni inverz matrice **Y**, obeležili bismo oznakom $\mathbf{Y}^{-1}$.

Regularni inverz mogu imati samo neke kvadratne matrice. Kvadratne matrice čija je determinanta jednaka nuli zovu se singularne matrice i nemaju regularni inverz. Za pravougaone matrice i singularne kvadratne matrice koje nemaju regularni inverz postoje tzv. **generalizovani inverzi**.

Sam postupak nalaženja regularnog inverza matrice prilično je zametan i neće biti objašnjen u ovom tekstu. Definisaćemo regularni inverz matrice preko njegovog ključnog svojstva:

Regularni inverz matrice **X**, u oznaci, $\mathbf{X}^{-1}$, je matrica koja ima sledeću osobinu:

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}\,\mathbf{X}^{-1} = \mathbf{I}$$

pri čemu je **I** matrica identiteta.

Dakle, kada matricu **X** pomnožimo njenim regularnim inverzom bilo sa koje strane dobijamo kao rezultat matricu identiteta, tj. dijagonalnu matricu u čijoj glavnoj dijagonali su svi elementi jednaki jedinici. Analogija sa "recipročnom vrednošću" nekog broja iz skalarne algebre je očigledna: kao što znamo iz skalarne algebre, kada se neki broj pomnoži svojom recipročnom vrednošću sa bilo koje strane, kao rezultat. množenja dobija se jedinica.

Zapamtite:

- $(\mathbf{X}^{-1})^{-1} = \mathbf{X}$. Regularni inverz inverza matrice **X** jednak je matrici **X**;
- $(\mathbf{XY})^{-1} = \mathbf{Y}^{-1}\,\mathbf{X}^{-1}$; $(\mathbf{X}_1\,\mathbf{X}_2 \cdots \mathbf{X}_k)^{-1} = \mathbf{X}_k^{-1} \cdots \mathbf{X}_2^{-1}\,\mathbf{X}_1^{-1}$.
  Regularni inverz proizvoda dveju ili više matrica jednak je proizvodu pojedinih regularnih inverza ovih matrica u obrnutom redosledu;
- Regularni inverz dijagonalne matrice **D** je dijagonalna matrica $\mathbf{D}^{-1}$ čiji su elementi na glavnoj dijagonali jednaki recipročnim vrednostima matrice **D**;
- Regularni inverz matrice identiteta **I** je matrica identiteta.

8. ORTOGONALNE MATRICE

Matrice u čijim kolonama ili redovima su vektori jedinične dužine, pri čemu bilo koja dva različita reda ili bilo koje dve različite kolone predstavljaju međusobno ortogonalne vektore, jesu ortogonalne matrice.

Preciznije, kvadratna matrica **P** je ortogonalna, akko ("ako i samo ako") je njen regularni inverz jednak njenom transponu, tj. akko $\mathbf{P}^{-1} = \mathbf{P}^{t}$.

Za ortogonalne matrice, stoga, važi i sledeće:

$$\mathbf{P}^{t}\mathbf{P} = \mathbf{P}\mathbf{P}^{t} = \mathbf{I}$$

9. RANG MATRICE

Rang matrice jednak je maksimalnom broju linearno nezavisnih redova ili kolona u matrici. Ako se radi o kvadratnoj matrici, njen rang može biti jednak broju redova ili kolona u matrici, ali može biti i manji. Ako je matrica pravougaona, njen rang ne može biti veći od njene manje dimenzije.

10. SVOJSTVENE VREDNOSTI I SVOJSTVENI VEKTORI MATRICE

Ako je **R** kvadratna matrica reda m x m, **x** vektor reda m x 1 čiji je bar jedan element različit od nule ($\mathbf{x} \neq \mathbf{0}$), a λ skalar, takav da je $\mathbf{Rx} = \lambda\mathbf{x}$, tada je λ svojstvena ili karakteristična vrednost (engl. eigenvalue, čita se "ajgenvelju") matrice **R**, a **x** je karakteristični ili svojstveni vektor matrice **R**. Svojstvene vrednosti i svojstveni vektori simetričnih matrica imaju neka važna i specifična svojstva koja nemaju svojstvene vrednosti i svojstveni vektori nesimetričnih kvadratnih matrica. Na određivanju svojstvenih vrednosti i svojstvenih vektora matrice zasniva se veliki broj multivarijacionih statističkih postupaka[9].

Matrična jednačina $\mathbf{Rx} = \lambda\mathbf{x}$, koja definiše svojstvene vrednosti može se, prebacivanjem izraza $\lambda\mathbf{x}$ na levu stranu napisati i u sledećem obliku: $\mathbf{Rx} - \lambda\mathbf{Ix} = \mathbf{0}$, pri čemu je **0** vektor nula, tj. vektor čiji su svi elementi jednaki nuli, a **I** matrica identiteta. Budući da kvadratna matrica reda m x m ima m svojstvenih vrednosti, onda ćemo uvesti indeks p uz oznaku svojstvene vrednosti i ovu matričnu jednačinu napisati u sledećem obliku: $\mathbf{Rx} - \lambda_p\mathbf{Ix} = \mathbf{0}$ Rešenje matrične jednačine $(\mathbf{R} - \lambda_p\mathbf{I})\,\mathbf{x}_p = \mathbf{0}$ se ne može dobiti, što je inače uobičajen postupak, njenim množenjem sa inverzom matrice $(\mathbf{R} - \lambda_p\mathbf{I})$, tj. matricom $(\mathbf{R}-\lambda_p\mathbf{I})^{-1}$, jer kada bismo to uradili dobili bismo $\mathbf{x}_p = (\mathbf{R} - \lambda_p\mathbf{I})^{-1}\,\mathbf{0} = \mathbf{0}$ što je u kontradikciji sa zahtevom da je $\mathbf{x}_p \neq \mathbf{0}$. Zato $(\mathbf{R}-\lambda_p\mathbf{I})$ ne sme imati regularni inverz, tj. mora biti singularna. To znači da tražimo λ_p takvo da je determinanta $|\mathbf{R}-\lambda_p\mathbf{I}| = 0$, a time se rešavanje svodi na rešavanje sistema m homogenih jednačina sa m nepoznatih. Dakle, svojstvene vrednosti, kada se oduzmu od dijagonalnih elemenata matrice **R** pretvaraju tu matricu u singularnu matricu, tj. matricu $\mathbf{R}- \lambda\mathbf{I}$ koja nema regularni inverz, tj. čija je determinanta $|\mathbf{R} - \lambda\mathbf{I}|$ jednaka nuli. Izraz $|\mathbf{R}-\lambda_p\mathbf{I}| = 0$ zove se karakteristična jednačina matrice **R**, a sama determinanta $|\mathbf{R} - \lambda\mathbf{I}|$ je karakteristična funkcija matrice **R**.

Daljim razvojem determinante dobijamo polinomsku jednačinu po λ stepena m:

$$f(\lambda) = (-\lambda)^m + b_{m-1}\,(-\lambda)^{m-1} + \ldots + b_1\,(-\lambda) + b_0 = 0$$

Ova jednačina je **karakteristični polinom** matrice **R**, a $\lambda_1, \lambda_2, \ldots, \lambda_m$ su karakteristični koreni ili **svojstvene vrednosti** (engl. eigenvalues) matrice **R**. Njihovim nalaženjem dobijamo i njima odgovarajuće karakteristične, svojstvene vektore $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. Dakle, svakoj svojstvenoj vrednosti odgovara određeni svojstveni vektor.

Ako konstruišemo dijagonalnu matricu

$$\Lambda = (\lambda_p) \quad (p = 1, 2, \ldots, m)$$

dakle, matricu čiji su dijagonalni elementi svojstvene vrednosti, poređane po veličini od gornjeg levog ka donjem desnom uglu matrice, i kvadratnu matricu

---

[9] Dodatna objašnjenja svojstvenih vrednosti i svojstvenih vektora data su u okviru izlaganja Analize glavnih komponenti (Glava XVIII Priručnika)

$$\mathbf{X} = (x_p) \quad (p = 1, 2, \ldots, m)$$

u čijim kolonama su svojstveni vektori tada, ako je matrica **R**, simetrična važi sledeće:

$$\mathbf{X}^t\mathbf{R}\mathbf{X} = \Lambda, \text{ a odatle } \mathbf{X}\Lambda\mathbf{X}^t = \mathbf{R}$$

Budući da je Λ dijagonalna matrica, očito je da se simetrična matrica može transformisati u dijagonalnu matricu njenim množenjem sa desne i leve strane ortogonalnom matricom i njenim transponom. Budući da se skup različitih svojstvenih vrednosti matrice zove spektralni skup, dekompozicija simetrične matrice na proizvod transpona ortogonalne matrice svojstvenih vektora, dijagonalne matrice svojstvenih vrednosti i ortogonalne matrice svojstvenih vektora naziva se spektralnom dekompozicijom matrice.

Zapamtite:

- Zbir svojstvenih vrednosti matrice **R** jednak je njenom tragu , a proizvod svojstvenih vrednosti matrice **R** jednak je determinantni te matrice:

$$\sum_{p=1}^{m}\lambda_p = \text{tr}(\mathbf{R}); \qquad \prod_{p=1}^{m}\lambda_p = |\mathbf{R}|$$

  Na osnovu drugog izraza očigledno je da, ako je matrica **R** singularna, bar jedna njena svojstvena vrednost mora biti jednaka 0 jer je determinanta singularne matrice jednaka nuli;

- Ako je matrica **R** simetrična, svojstveni vektori koji odgovaraju različitim svojstvenim vrednostima ove matrice su međusobno ortogonalni. Ako, pak, matrica **R** nije simetrična, svojstveni vektori koji odgovaraju različitim svojstvenim vrednostima ove matrice nisu međusobno ortogonalni.

11. KVADRATNA FORMA

Ako je **R** simetrična matrica reda m x m, a **b** vektor reda m x 1, tada se izraz

$$\mathbf{b}^t\mathbf{R}\mathbf{b} = \sum_{k=1}^{m}\sum_{j=1}^{m} b_j b_k r_{jk}$$

zove kvadratna forma matrice **R**. Na osnovu pravila o množenju matrica očigledno je da se kao rezultat množenja $\mathbf{b}^t\mathbf{R}\mathbf{b}$ dobija skalar, tj. jedan broj.

Kvadratnom formom matrice u statistici se često izražava varijansa linearne kombinacije varijabli koje su u kolonama matrice.

DODATAK 4: ISPRAVKE KRUPNIJIH GREŠAKA U PRIRUČNIKU I VAŽNE NAPOMENE I DOPUNE U VEZI SA POJEDINIM DELOVIMA OSNOVNOG TEKSTA

SADRŽAJ DODATKA 4

A. ISPRAVKE KRUPNIJIH GREŠAKA U PRVOM IZDANJU PRIRUČNIKA 2

B. NAPOMENE I DOPUNE U VEZI SA POJEDINIM DELOVIMA OSNOVNOG TEKSTA PRIRUČNIKA 6

I. UVOD U PROGRAMSKI PAKET SPSS for WINDOWS (napomena 1) 6

II. ORGANIZACIJA I UNOS PODATAKA (napomena 2) 6

III. OSNOVNI POJMOVI I OZNAKE ČIJE POZNAVANJE JE NEOPHODNO ZA RAZUMEVANJE DALJEG TEKSTA (napomena 3) 7

Operator Σ -operator sabiranja ili sumacioni operator (dopuna 4) 7

Operator ΣΣ -operator dvostrukog sabiranja ili dvojni sumacioni operator (dopuna 5) 9

Operator proizvoda Π (dopuna 6) 10

IV. SREĐIVANJE PODATAKA - STATISTIČKI OPIS (I GRAFIČKI PRIKAZ) POPULACIJE ILI UZORKA U POGLEDU JEDNE KVANTITATIVNE VARIJABLE (napomena 7) 11

Značenje dugmadi u prozorima za dijalog komande Compute (dopuna 8) 11

IV.8. Neke važne transformacije podataka na kvantitativnoj varijabli (dopuna 9) 12

V. STATISTIČKI OPIS I GRAFIČKI PRIKAZ JEDNE KATEGORIČKE (NOMINALNE) VARIJABLE (dopuna 10) 17

Entropija nominalne varijable (dopuna 10) 17

Logaritamska funkcija (dopuna 10) 17

Osnovne osobine logaritama (dopuna 10) 17

XIV. MULTIPLA REGRESIONA ANALIZA (dopuna 11) 18

XIV.4. Medijatorske, moderatorske i supresorske prediktorske varijable u regresionoj analizi (dopuna 11) 18

A. ISPRAVKE KRUPNIJIH GREŠAKA U PRVOM IZDANJU PRIRUČNIKA

U Tabeli D4.1 ovog dodatka date su samo najvažnije ispravke grešaka u prvom izdanju Priručnika koje je autor otkrio do sada. "Providne" pravopisne greške nisu unete u tabelu jer se pretpostavlja da ih čitalac može sam otkriti i sa sigurnošću pogoditi ispravnu reč. Takve ispravke ne bi donele mnogo koristi čitaocu, a tabela bi postala znatno složenija.

Posle tabele sa ispravkama date su pojedine važne napomene koje bi trebalo da čitaocu olakšaju praćenje pojedinih delova Priručnika. Napomene i dopune su grupisane prema poglavljima osnovnog teksta Priručnika na koja se odnose. Svaka napomena ima redni broj pored kojeg u zagradi stoje orijentiri na osnovu kojih čitalac može pronaći na koji deo teksta u Priručniku se napomena ili dopuna odnosi.

Tabela D4.1: Ispravke krupnijih grešaka u Priručniku

Strana	Red	Stoji	Treba da stoji
1	Fusnota1, Treći red	View Options	Edit Options
10	Fusnota 7, Sedmi red	Dva skupa su identično jednaka	Dva skupa su, ako se radi o tzv. neuređenim skupovima, identično jednaka
12	Fusnota13, Prvi red	Viks	Vilks
13	Fusnota 15		Osnovni pojmovi matrične algebre prikazani su u veoma sažetom vidu u Dodatku 3 ovog Priručnika
15	Osmi red odozdo	Nemaju realne nule pa množenje konstantom nema smisla.	Nemaju realne nule, a budući da je jedinica merenja prirodna (nearbitrarna) množenje konstantom nema smisla.
17	Sedmi red odozgo	imati (pored relacije ekvivalencije i potpunog poretka postoji specifikovan racio bilo koja	imati (pored relacije ekvivalencije i potpunog poretka) i specifikovan racio bilo koja
26	Prvi red odozdo	$\sum_{i=1}^{n}(x_i - x_0) \mid x_0 \neq M_x$	$\sum_{i=1}^{n}(x_i - x_0)^2 \mid x_0 \neq M_x$
26	Fusnota 22, Drugi red	Elementi linearne algebre prikazani su u dodatku knjizi *Teorija merenja I* K.Momirovića, B. Wolfa i D. A Popovića.	Osnovni pojmovi iz matrične algebre dati su veoma sažeto u Dodatku 3, a kompletniji prikaz linearne algebre može se naći u dodatku knjizi *Teorija merenja I* K.Momirovića, B. Wolfa i D. A Popovića.

31	Šesti red odozdo	$\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij} / (n^2 - n)$	$\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij} / (n^2 - n)$ (Operator dvojnog sabiranja objašnjen je u Dodatku 4 - napomena broj 5).
56	Dvadeset drugi red odozgo	Iskaz ¬ B	Iskaz ¬ A
60	Sedmi red odozdo	$SE_{Sk} = \sqrt{\frac{24}{n}}$	$SE_{Ku} = \sqrt{\frac{24}{n}}$
62	Treći red odozgo	Šapiro-Viksov statistik	Šapiro-Vilksov statistik
88	Fusnota 62, Prvi red	(fusnota 41)	(fusnota 52)
92	Jedanaesti red odozgo	sa (r - 1) (k - 1) i (n - rk) stepeni	sa (s - 1) (g - 1) i (n - s*g) stepeni
92	Deveti red odozdo	sa (r - 1) i (n - n_k) stepeni slobode	sa (s - 1) i (n - s*g) stepeni slobode
93	Treći red odozgo	sa (k - 1) stepeni slobode.	sa (g - 1) i (n - s*g) stepeni slobode.
103	Sedmi red odozgo	kriterijumske varijable y	kriterijumske varijable u
105	Treći red odozgo	ili skalarni proizvod vektora u kojima su rezultati na varijablama x i y iskazani u standardizovanom obliku:	ili količnik skalarnog proizvoda vektora u kojima su rezultati na varijablama x i y iskazani u standardizovanom obliku i broja parova rezultata, tj. veličine uzorka
106	Dvadeset treći red odozgo	u Odeljku **VI.2.**	u Odeljku **IX**
106	Fusnota 68, Prvi red	pri čemu je cos θ = $r_{xy} * r_{xz}$	pri čemu je cos θ = $r_{xy} = r_{xz}$
114	Jedanaesti red odozgo	$\mathbf{z_x b^* = z_y + g} \mid \sigma^2_g = \mathbf{g^t g} * 1/n = \text{minimum}$	$\mathbf{z_x b^* = z_y + g} \mid \sigma^2_g = \mathbf{g^t g}\, n^{-1} = \text{minimum}$
116	Petnaesti red odozgo	$\sigma_g^2 = \sigma_y^2 = \rho^2\sigma_y^2 = \sigma_y^2(1-\rho^2)$	$\sigma_g^2 = \sigma_y^2 - \rho^2\sigma_y^2 = \sigma_y^2(1-\rho^2)$
117	Šesti red odozdo	$t = (a - \alpha) / SE_{\alpha,,}$	$t = (a - \alpha) / SE_a$,
123	Prvi red odozdo	sa k-1 stepeni slobode	sa r-1 stepeni slobode
124	Četrnaesti red odozdo	$\phi_k = n/k$	$\phi_k = n/r$
126	Jedanaesti red odozgo	f_{jk}=numerus($e_i \in q_j \wedge e_i \in q_k$)	f_{jk}= numerus ($e_i \in q_{1j} \wedge e_i \in q_{2k}$)

126	Dvanaesti red odozgo	$[p_{jk} = P(e_i \in q_j \cap q_k)]$	$[p_{jk} = P(e_i \in q_{1j} \wedge q_{2k})]$
126	Jedanaesti red odozdo	H_o: $p_{jk} = p_j * p_k$; $\forall j$ i $\forall k$	H_o: $p_{jk} = p_{1j} * p_{2k}$; $\forall j$ i $\forall k$
126	Treći red odozdo	H_1: $p_{jk} \neq p_j * p_k$	H_1: $p_{jk} \neq p_{1j} * p_{2k}$
127	Fusnota 81, Peti red	$P[(Q_1=q_j) \wedge (Q_2 = q_k)] = P(Q_1=q_j) * P(Q_2 = q_k)$.	$P[(Q_1=q_{1j}) \wedge (Q_2 = q_{2k})] = P(Q_1=q_{1j}) * P(Q_2 = q_{2k})$.
127	Fusnota 81, Šesti red	Tada $p_{jk} = p_j * p_k = (f_j/n) * (f_k/n)$, a $\phi_{jk} = p_{jk} * n = (f_j/n) * (f_k/n) * n = (f_j * f_k)/n$.	Tada $p_{jk} = p_{1j} * p_{2k} = (f_{1j}/n) * (f_{2k}/n)$, a $\phi_{jk} = p_{jk} * n = (f_{1j}/n) * (f_{2k}/n) * n = (f_{1j} * f_{2k})/n$.
128	Treći red odozdo	$C_{max} = \sqrt{q - 1 / q}$	$C_{max} = \sqrt{(q-1)/q}$
143	Četvrti red odozgo	$z_{dobijeno} = (U - \mu_U) / \sigma_U$	$z = (U - \mu_U) / \sigma_U$
151	Dvanaesti red odozgo	u = kvantitativna kriterijumska varijabla.	u je kvantitativna kriterijumska varijabla.
155	Trinaesti red odozgo	$f = \frac{R^2_{y.1,2,..m}/m}{1-R^2_{y.1,2,..m}/n-m-1}$	$f = \frac{R^2_{y.1,2,..m}/m}{(1-R^2_{y.1,2,..m})/(n-m-1)}$
155	Trianesti red odozgo	$f_j = r^2_{y.xj}(n-2)/ 1- r^2_{y.xj}$	$f_j = r^2_{y.xj}(n-2)/ (1- r^2_{y.xj})$
156	Fusnota 97	u odeljku **XIV. 2.**	u odeljku **XIV. 3.**
157	Fusnota 99, Drugi red	u odeljku **XIV.1.** ovog teksta	u odeljku **XIV.3.** ovog teksta
171	Osamnaesti red odozgo	c) $s(e_r, e_s) = 0 \mid e_r = e_s$.	c) $s(e_r, e_s) = 1 \mid e_r = e_s$.
189	Drugi red odozgo	1. $\mathbf{k}_{1p}^t \mathbf{k}_{2p} n^{-1} = \mathbf{x}_{1p}^t \mathbf{R}_{12} \mathbf{x}_{2p} n^{-1}$ = maximum	1. $\mathbf{k}_{1p}^t \mathbf{k}_{2p} n^{-1} = \mathbf{x}_{1p}^t \mathbf{R}_{12} \mathbf{x}_{2p}$ = maximum
189	Deveti red odozdo	$\mathbf{x}_{1p}^t \mathbf{R}_{12} \mathbf{x}_{2p} n^{-1} - (1/2)\lambda_{1p}(1- \mathbf{x}_{1p}^t \mathbf{R}_{11} \mathbf{x}_{1p} n^{-1}) - (1/2)\lambda_{2p}(1- \mathbf{x}_{2p}^t \mathbf{R}_{22} \mathbf{x}_{2p} n^{-1})$	$\mathbf{x}_{1p}^t \mathbf{R}_{12} \mathbf{x}_{2p} - (1/2)\lambda_{1p}(1- \mathbf{x}_{1p}^t \mathbf{R}_{11} \mathbf{x}_{1p}) - (1/2)\lambda_{2p}(1- \mathbf{x}_{2p}^t \mathbf{R}_{22} \mathbf{x}_{2p})$
190	Dvadeset prvi red odozgo	$\chi^2_p = - [n-1/2 (m_1 + m_2 + 3)] \ln \Lambda_p$	$\chi^2_p = -[n - (\frac{m_1 + m_2 + 3}{2})] \ln \Lambda_p$
200	Osmi red odozdo	$V = \{ v_j ; j = 1,2,\ldots, m\} \subset U$ skup kvantitativnih varijabli	$V = \{ v_j ; j = 1,2,\ldots, m\} \subset U$ skup kvantitativnih varijabli
204	Prvi red odozdo	$\chi^2_p = - [n-1/2 (m+g+3)] \ln \Lambda_p$,	$\chi^2_p = -[n - (\frac{m+g+3}{2})] \ln \Lambda_p$

220	Osamnaesti red odozgo	$V = \{ v_j ; j = 1,2,\ldots, m\} \subset U$ je skup kvantitativnih varijabli koji reprezentuje univerzum U	$V = \{ v_j ; j = 1,2,\ldots, m\} \subset U$ je skup kvantitativnih varijabli koji reprezentuje univerzum U
221	Dvadeset deveti red odozgo	$\sigma_p^2 = \mathbf{k}_p^t \mathbf{k}_p = \mathbf{x}_p^t \mathbf{R} \mathbf{x}_p$ = maximum	$\sigma_p^2 = \mathbf{k}_p^t \mathbf{k}_p n^{-1} = \mathbf{x}_p^t \mathbf{R} \mathbf{x}_p$ = maximum
223	Trinaesti red odozgo	$(\forall w \in P)$	$(\forall w \in \mathcal{R}$, pri čemu je $\mathcal{R}$ skup realnih brojeva)
223	Sedmi red odozdo	$\mathbf{K}^t\mathbf{K} = \mathbf{X}^t\mathbf{Z}^t\mathbf{Z}\mathbf{X} = \mathbf{X}^t\mathbf{R}\mathbf{X} = \Lambda$.	$\mathbf{K}^t\mathbf{K}\, n^{-1} = \mathbf{X}^t\mathbf{Z}^t\mathbf{Z}\mathbf{X}\, n^{-1} = \mathbf{X}^t\mathbf{R}\mathbf{X} = \Lambda$.
223	Prvi red odozdo	$\mathbf{F} = \mathbf{Z}^t\mathbf{K}_s = \mathbf{R}\mathbf{X}\Lambda^{-1/2} =$	$\mathbf{F} = \mathbf{Z}^t\mathbf{K}_s\, n^{-1} = \mathbf{R}\mathbf{X}\Lambda^{-1/2} =$
223	Fusnota 113, Drugi red	$\mathbf{x}_q = \lambda_p \mathbf{x}_p^t \mathbf{x}_{pq}$	$\mathbf{x}_q = \lambda_p \mathbf{x}_p^t \mathbf{x}_p$
224	Prvi red odozgo	$\mathbf{F} = \mathbf{Z}^t\mathbf{K}_s = \mathbf{X}\Lambda^{-1/2}$	$\mathbf{F} = \mathbf{Z}^t\mathbf{K}_s\, n^{-1} = \mathbf{X}\Lambda^{-1/2}$
Dodata k2, strana 15	Trinaesti red odozgo	$\sigma^2_{A+B1X1+\ldots+BmXm} = B_1^2\sigma^2_{X1} + \ldots + B_m^2\sigma^2_{Xm} + 2\, B_1\, B_2\, \sigma_{X1X2} + \ldots + 2\, B_{m-1}\, B_m\, \sigma^2_{Xm-1\, Xm}$.	$\sigma^2_{A+B1X1+\ldots+BmXm} = B_1^2\sigma^2_{X1} + \ldots + B_m^2\sigma^2_{Xm} + 2\, B_1\, B_2\, \sigma_{X1X2} + \ldots + 2\, B_{m-1}\, B_m\, \sigma_{Xm-1\, Xm}$.
Dodata k2, strana 16	Deseti red odozgo	Uniformna distribucija	Uniformna distribucija (za diskretnu slučajnu varijablu)
Dodata k2, strana 17	Prvi red odozdo	$\sigma^2 = (1-p)/[(1-p)-1]$	$\sigma^2 = (1-p)/p^2$
Dodata k2, strana 18	Četvrti red odozdo	Uniformna (rektagularna) funkcija gustine	Uniformna (rektagularna) funkcija gustine (za kontinuiranu slučajnu varijablu)
Dodata k2, strana 19	Petnaesti red odozdo	68.26 ukupne površine	68.26 % ukupne površine
Dodata k2, strana 24	Sedmi red odozdo	Empiriska distribucija učestalosti	Empirijska distribucija učestalosti

B. NAPOMENE I DOPUNE U VEZI SA POJEDINIM DELOVIMA OSNOVNOG TEKSTA PRIRUČNIKA

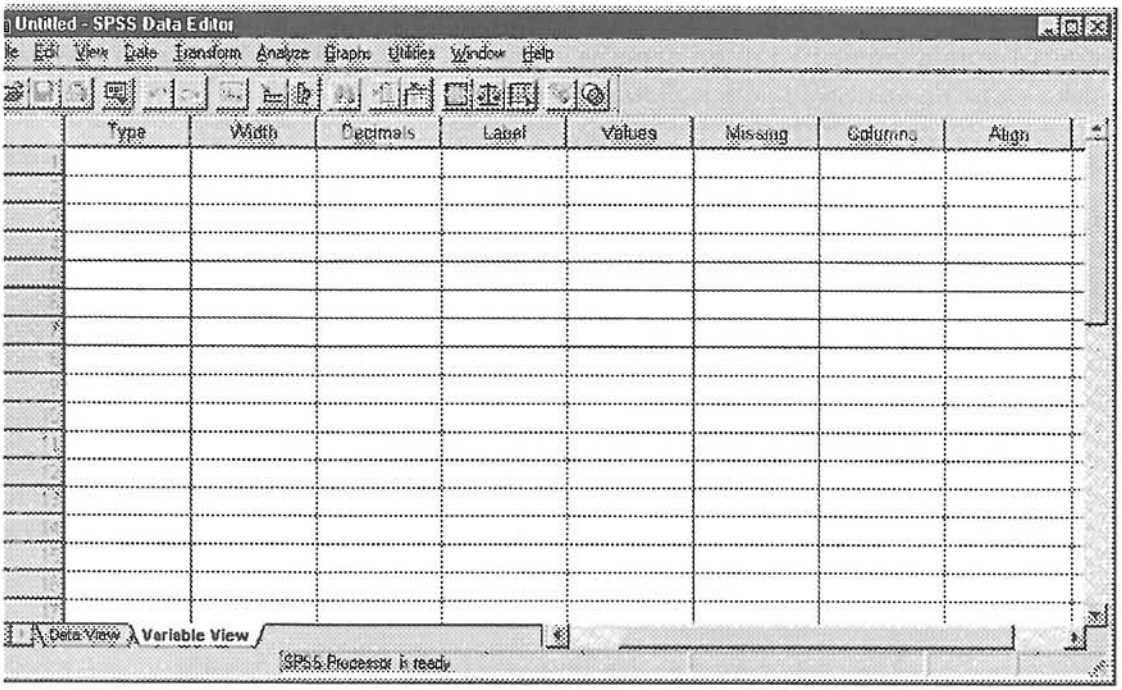
I. UVOD U PROGRAMSKI PAKET SPSS for WINDOWS

1. (strana 2, tekst koji se odnosi na deo **3. Prozor za podatke**) U verzijama SPSS-a počevši od verzije 10.0 prozor editora podataka promenjen je tako da se mogu prema potrebi posmatrati sami podaci (**Data view** opcija) ili varijable (**Variable view** opcija), tj. imena (**name**) i sve druge karakteristike varijabli (tip- **Type**, naziv varijable-**Label**, oznake vrednosti (**Values**) itd. Prelazak iz jedne u drugu opciju moguć je bilo klikom miša na odgovarajući jezičak u levom donjem uglu prozora editora podataka, bilo odabirom opcije **Variables** (kada je trenutno aktivna opcija **Data View**) ili opcije **Data** (kada je trenutno aktivna opcija **Data View**) u okviru stavke **View** osnovnog menija programa na vrhu prozora za podatke.

II. ORGANIZACIJA I UNOS PODATAKA

2. (strana 5, odeljak II.1. Definisanje varijabli) U verzijama SPSS-a počevši od verzije 10.0 radi definisanja varijabli potrebno je kliknuti na jezičak **Variable view** na dnu prozora za podatke (slika D4.1). Zatim se u polje **Name** upisuje ime varijable, a u polje **Label** naziv varijable (ako je potrebno). U poddijaloge za definisanje tipa (**Type**), naziva vrednosti (polje **Values**) i korisničkih nedostajućih vrednosti (polje **Missing**) ulazi se tako što se klikne na odgovarajuće polje i zatim na desni deo okvira na kojem se nalaze tri tačke (slika D4.1, polje Type). Definisanje ostalih karakteristika varijable: širine (**Width**), broja decimalnih mesta (**Decimals**), formata ispisa varijable (**Columns**), poravnanja podataka u odnosu na levu i desnu marginu ćelije (**Align**) i merne skala (**Measure**) izvodi se tako što se klikne u odgovarajuću ćeliju, a zatim na strelicu na desnoj strani ćelije koja se pri tom pojavljuje.
   Ukoliko se prilikom definisanja varijabli u verziji 10.0 SPSS-a primeni postupak koji je važio u prethodnim verzijama programa (dvostruki klik mišem na naslov kolone var0000m) automatski će biti aktiviran prozor **Variable view**.

Slika D4.1: Izgled prozora **Variable view** editora podataka (SPSS 10.0 i novije verzije)



III. OSNOVNI POJMOVI I OZNAKE ČIJE POZNAVANJE JE NEOPHODNO ZA RAZUMEVANJE DALJEG TEKSTA

3. (strana 14, uz definiciju pojma **Merenje**). Osnovne operacije koje se primenjuju u bilo kojem psihološkom merenju su: 1. brojanje; 2. uređivanje u redosled (rangovanje) i 3. razvrstavanje (klasifikovanje). Svi postupci merenja mogu se svesti na ove tri bazične operacije. Pri merenju visine ili inteligencije, na primer, broji se koliko jedinica ima neki ispitanik (santimetara za visinu ili broja tačno urađenih zadataka na testu inteligencije). Na osnovu odgovora na pitanje o polu ili nacionalnosti ispitanik se klasifikuje u jednu od kategorija varijabli pol ili nacionalnost. Isto tako, određeni procenjivač (ili određeni kriterijum) može poslužiti za rangovanje ispitanika prema nekoj osobini (na primer hrabrosti, urednosti). Koliko ispitanika u nekom uzorku pripada određenoj kategoriji neke varijable (na primer grupi ljudi sa osnovnim ili visokim obrazovanjem) možemo dobiti prostim prebrojavanjem odgovora ispitanika na pitanje o školskoj spremi.

4. (osnovni pojam neophodan za dalje praćenje teksta)

Operator Σ -operator sabiranja ili sumacioni operator

Operatorom sabiranja, u oznaci Σ, može se sažeto prikazati sabiranje indeksiranih vrednosti varijabli, tj. elemenata koji se označavaju indeksiranim opštim članom (indeksiranom **promenljivom**) koji se nalazi posle operatora. Izraz

$\sum_{i=1}^{n} x_i$ čitamo kao "Suma x i, za i od 1 do n", a elementi ovog izraza su znak sume (Σ), promenljiva, opšti član ili **sumand** (x) i indeks sabiranja (i). Indeks sabiranja ima donju i gornju granicu koje su napisane ispod i iznad znaka Σ. U statistici su najčešće donja i gornja granica 1 i n, pri čemu je 1 prvi a n poslednji rezultat u nizu koje ima neka varijabla na uzorku od n ispitanika ili entiteta.

Tako, izraz $\sum_{i=1}^{n} x_i$, $i = 1, 2, ..., n$ znači da se sabiraju svi elementi nekog indeksiranog niza od prvog do n-tog elementa:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 \cdots x_n .$$

Ako želimo, na primer, da saberemo rezultate pet ispitanika redom, pri čemu je rezultat prvog 12, drugog 10, trećeg 8, četvrtog 10 i petog ispitanika 5, onda opšti član x označava bilo koji rezultat, a na redosled ispitanika ili rezultata koji se sabira upućuje indeks uz opšti član. U ovom slučaju operator sabiranja bi sažeto označavao sledeći postupak:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 12 + 10 + 8 + 10 + 5 = 45 .$$

Ukoliko se podrazumeva da se svi elementi skupa označeni opštim članom sabiraju onda se indeks u operatoru ponekad izostavlja. U prethodnom primeru, ako je jasno da se sabira svih n elemenata, tj. svih 5 rezultata mogli bismo operator sabiranja napisati i ovako:

$$\sum x = x_1 + x_2 + x_3 + x_4 + x_5 = 12 + 10 + 8 + 10 + 5 = 45 .$$

Veoma je korisno za razumevanje obrazaca u ovom Priručniku, kao i za praćenje literature iz oblasti statistike naučiti nekoliko osnovnih pravila koja važe za ovaj operator:

- Ako je $x_i = k$ = konstanta za svako i, tada je

$$\sum_{i=1}^{n} x_i = n * k$$

(Znak * je oznaka množenja).

Drugačije rečeno zbir od n istih elemenata (konstanti) jednak je prizvodu broja elemenata i konstante kojoj su jednaki svi elementi;

- Ako je k = konstanta, tada je

$$\sum_{i=1}^{n} k * x_i = k * \sum_{i=1}^{n} x_i .$$

Dakle zbir proizvoda konstante i promenljive jednak je proizvodu konstante i zbira promenljive;

- Ako su x_i i y_i opšti članovi dva niza (dve varijable ili dva indeksirana sumanda) pri čemu je u oba niza indeks i isti, onda

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Iskazano rečima, zbir za zbir opštih elemenata jednak je zbiru zbirova pojedinih opštih elemenata ili varijabli.

- Ako se nekoliko matematičkih operacija, uključujući i sabiranje, odvija na indeksiranom opštem članu, prvo se vrši stepenovanje, zatim deljenje i množenje, a sabiranje se obavlja poslednje, osim ako zagradama nije drugačije naznačeno. Na primer:

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2.$$

Ali, zbog zagrade, važi sledeće:

$$(\sum_{i=1}^{n} x_i)^2 = (x_1 + x_2 + \cdots + x_n)^2.$$

Zapamtite: $\sum_{i=1}^{n}(x_i * y_i) \neq \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i$.

Zbir proizvoda opštih elemenata (promenljivih) **nije jednak** proizvodu zbirova pojedinih opštih elemenata ili promenljivih.

5. (osnovni pojam neophodan za dalje praćenje teksta)

Operator $\Sigma\Sigma$ -operator dvostrukog sabiranja ili dvojni sumacioni operator

Ovaj operator se definiše na sledeći način:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = \sum_{i=1}^{n}(x_{i1} + x_{i2} + \cdots + x_{im}) = (x_{11} + x_{12} + \cdots + x_{1m}) + (x_{21} + x_{22} + \cdots + x_{2m}) + \cdots + (x_{n1} + x_{n2} + \cdots x_{nm})$$

Uočimo da varijabla ili opšti član imaju dva indeksa: i i j. Svaki od ovih indeksa ima donju i gornju granicu sabiranja. Potrebno je za svaku vrednost indeksa spoljašnje ("leve") sume napraviti zbir svih elemenata po indeksu j. Dakle, za prvu vrednost indeksa i (1) napravimo zbir svih elemenata za j od 1 do m i to isto ponovimo za svaku vrednost indeksa i. Na kraju saberemo sve ove zbirove i dobijemo zbir od ukupno n*m elemenata, pri čemu su n i m gornje granice sabiranja za indekse i i j. Na ovaj način može se, na primer, dobiti zbir svih elemenata neke matrice u kojoj i najčešće predstavlja indeks reda, a j indeks kolone u kojoj se nalazi element koji se sabira (o matricama se može pročitati u Dodatku 3).

Bitno svojstvo ovog operatora je zamenljivost mesta dveju suma:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij}$$

6. (osnovni pojam neophodan za dalje praćenje teksta)

Operator proizvoda $\prod$

Operator proizvoda, u oznaci $\prod_{i=1}^{n} x_i$ (čita se kao "proizvod x i, za i od 1 do n"), definiše se na sledeći način:

$$\prod_{i=1}^{n} x_i = x_1 * x_2 * \cdots * x_n .$$

Radi se očigledno o operatoru kojim se sažeto prikazuje množenje indeksiranih elemenata nekog niza. Ako su 5, 2, 1 i 3 elementi nekog niza koji su dati ovim redosledom onda je $x_1 = 5$, $x_2 = 2$, $x_3 = 1$, a $x_4 = 3$, pa je proizvod članova ovog niza za i od 1 do 4:

$$\prod_{i=1}^{4} x_i = x_1 * x_2 * x_3 * x_4 = 5*2*1*3 = 30$$

Osnovna pravila koja važe za ovaj operator su:

- Ako je $x_i = k$ = konstanta za svako i, tada je

$$\prod_{i=1}^{n} x_i = \prod_{i=1}^{n} k = k^n .$$

Dakle, proizvod konstanti za i od 1 do n jednak je n-tom stepenu konstante;

- Ako je k = konstanta, a x promenljiva, tada je

$$\prod_{i=1}^{n} kx_i = k\prod_{i=1}^{n} x_i .$$

Dakle proizvod za produkt konstante i promenljive jednak je proizvodu konstante i proizvoda promenljive;

- Ako su x_i i y_i opšti članovi dva niza (dve varijable) pri čemu je u oba niza indeks i isti, onda je

$$\prod_{i=1}^{n} x_i y_i = (\prod_{i=1}^{n} x_i)(\prod_{i=1}^{n} y_i)$$

Proizvod proizvoda dveju promenljivih jednak je proizvodu odvojenih proizvoda za svaku promenljivu;

- $\ln \prod_{i=1}^{n} x_i = \sum_{1}^{n} \ln x_i$.

  Logaritam proizvoda vrednosti promenljive jednak je zbiru logaritama vrednosti promenljive.

IV. SREĐIVANJE PODATAKA - STATISTIČKI OPIS (I GRAFIČKI PRIKAZ) POPULACIJE ILI UZORKA U POGLEDU JEDNE KVANTITATIVNE VARIJABLE

7. (strana 20. i nadalje) Stavka **Statistics** u osnovnom meniju programa SPSS koja je postojala do verzije 8.0 zamenjena je u novijim verzijama programa (10.0 pa nadalje) stavkom **Analyze.** Prema tome korisnik Priručnika koji radi u nekoj od novijih verzija SPSS-a treba da umesto stavke **Statistics** na koju se upućuje u Priručniku pri definisanju statističkih procedura odabere stavku **Analize** u osnovnom meniju programa. Isto tako, treba imati na umu i da stavki **Summarize**, koja je postojala u starijim verzijama programa, u novijim verzijama odgovara stavka **Descriptive Statistics**.

8. (vrh strane 24) Značenje dugmadi u prozorima za dijalog komande Compute
U prozoru za dijalog komande **COMPUTE**, kao i u zadavanju uslova (dugme **IF**) u okviru komandi **COUNT** i **RECODE**, nalazi se tastatura sa brojevima i ostalim oznakama koje mogu biti korišćene u definisanju ovih komandi (slikaD4.2). U Tabeli D3.1 data su objašnjenja samo onih oznaka za koje postoji mogućnost da korisniku nisu već poznate.

Slika D4.2: Izgled prozora za poddijalog **If Cases** komande **Compute**

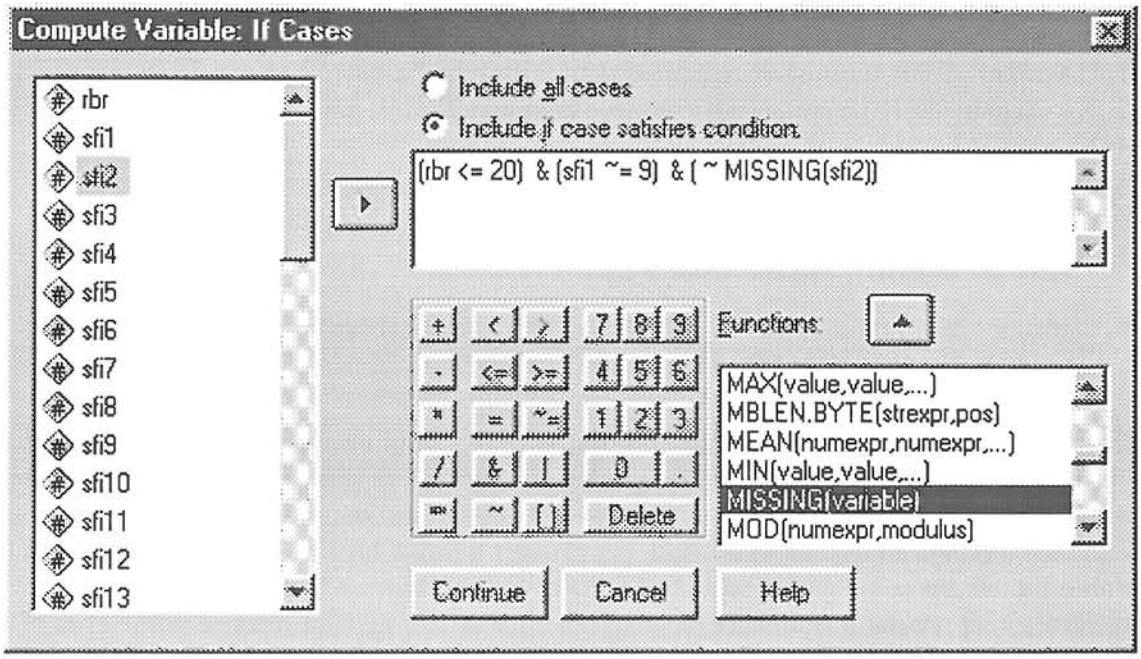


Tabela D4.2: Značenja pojedinih oznaka na dugmadima u poddijalozima komandi **COMPUTE, RECODE i COUNT.**

Oznaka na dugmetu	Značenje
≤	Manje ili jednako (u matematičkom značenju)
≥	Veće ili jednako (u matematičkom značenju)
*	Oznaka množenja: p1*p2 znači množenje varijabli čija su imena (engl. Name) p1 i p2.
~=	Nije jednako (u matematičkom značenju): p1 ~= 3 znači da vrednost na varijabli čije je ime p1 nije jednaka 3.
/	Oznaka deljenja: p1/p2 znači deljenje varijabli čija su imena p1 i p2.
&	Ovaj znak se zove "ampersen" i označava logičko I, tj. konjukciju. Tako, (p1 = 2) & (p2 ~= 9) pri zadavanju nekog uslova znači da oba uslova moraju biti ispunjena.
|	Označava logičko ILI, tj. disjunkciju. Tako, (p1 = 2) | (p2 ~= 9) pri zadavanju nekog uslova znači da bar jedan uslov (p1 = 2) ili (p2 ~= 9) mora biti ispunjen.
**	Oznaka stepenovanja: p1**2 znači da rezultate na varijabli čije je ime p1 treba kvadrirati
~	Ovaj znak se zove "tilda" i označava logičku negaciju. Tako izrazom ~(MISSING(p1)) postavljamo uslov da na varijabli p1 nema nedostajućih vrednosti (funkcija MISSING)
( )	Treba uočiti da klik na dugme sa oznakom otvorene i zatvorene zagrade daje obe zagrade odjednom.

Kombinovanjem pojedinih od ovih oznaka mogu se definisati, između ostalog, uslovi koje podaci treba da ispune da bi se na njima primenila određena procedura (matematička ili statistička). U primeru koji je prikazan na slici D4.2 postavljena su tri uslova koje ispitanik (čiji podaci su u aktivnom fajlu) treba da ispuni da bi se transformacija koja je prethodno definisana u polju **Numeric Expression** komande **Compute** primenila na njega:

1. U varijabli RBR mora imati vrednost manju ili jednaku 20;
2. Na varijabli SFI1 ne sme imati vrednost vrednost 9, i
3. Na varijabli SFI2 ne sme imati nedostajuću vrednost.

Samo na onim ispitanicima koji ispunjavaju sva tri uslova (uslovi u primeru povezani su znakom konjukcije) primeniće se transformacija koja je zadata u polju **Numeric Expression** komande **Compute**.

9. (strana 35, odeljak **IV.8. Neke važne transformacije podataka na kvantitativnoj varijabli**)

Sirovi ili izvorni rezultati na kvantitativnoj varijabli, koji se dobijaju kao neka vrsta "mere" ispitanika u pogledu određenog obeležja moraju se ponekada transformisati u drugačiji oblik. U psihologiji je veoma čest slučaj da tek transformisani izvorni rezultati merenja mogu dobiti smisao. Na primer, ukupan broj poena koje je ispitanik dobio na nekom testu sposobnosti često se prevodi u **standardizovani oblik** (z-skor), tj. od sirovog rezultata ispitanika na testu oduzima se aritmetička sredina grupe koja je testirana, a zatim se tako dobijena **devijaciona mera** deli standardnom devijacijom

cele grupe. Isto tako, prevodjenjem u percentilne rangove ili, pak, u T-vrednosti (iskazivanjem na skali čija je aritmetička sredina 50, a standardna devijacija 10) rezultati ispitanikovog postignuća na nekom psihološkom mernom instrumentu dobijaju jasan smisao koji se ne bi mogao videti iz sirovog broja postignutih poena na testu ili upitniku.

P. Horst (Horst, 1966, str. 70) deli sve transformacije sirovih ili izvornih rezultata koje se koriste u psihologiji u dve osnovne grupe:

1. Algebarske (ili kako ih Horst naziva "matematičke")
i
2. Površinske .

1. Algebarske transformacije se sastoje u tome da se, primenom odredjene matematičke, algebarske formule, tj. linearne ili nelinearne funkcije, sirove, izvorne vrednosti dobijene merenjem preslikaju u drugačiji oblik. Najpoznatiji primeri linearnih transformacija u psihologiji su centriranje, standardizacija kao i prevodjenje u T-skorove ili, pak, u C-skorove, a najčešće korišćene nelinearne transformacije u psihologiji su polinomske, logaritamske, logit, probit, eksponencijalne, kvadratno-korenske (engl. square root), recipročne i trigonometrijske (posebno arkussinusna).

Linearne transformacije se daleko češće koriste u analizi podataka u psihologiji od nelinearnih. Ove transformacije zasnivaju se na linearnom preslikavanju, tj. na linearnoj funkciji oblika

$$f(x) = a + bx$$

ili

$$y = a + bx$$

Radi se, zapravo, o tome da se, množenjem vrednosti na izvornoj varijabli x pomoću tzv. multiplikativne ili skalirajuće konstante (b) i dodavanjem na tako dobijeni rezultat tzv. aditivne konstante (a), dobija nova varijabla y.

U psihologiji je često potrebno izvorne rezultate linearno transformisati na skalu koja ima određenu vrednost aritmetičke sredine i standardne devijacije. Na primer, rezultati sa testa inteligencije ponekad se, radi lakše interpretacije i poređenja, prevode u skalu čija je aritmetička sredina 100, a standardna devijacija 10. Isto tako, veoma se često rezultati sa upitnika ličnosti prevode u skalu T-skorova, tj. u skalu čija je aritmetička sredina 50, a standardna devijacija 10. Za rešavanje problema transformisanja izvornih rezultata u takav oblik, kojim se postiže to da rezultati u novom, transformisanom obliku imaju odredjenu željenju vrednost aritmetičke sredine i standardne devijacije potrebno je odrediti samo vrednosti aditivne (a) i multiplikativne (b) konstante u linearnoj transformaciji. Jednostavan način da se ovaj problem reši u opštem slučaju je da se postave dve nezavisne jednačine koje govore o tome kako aditivna i multiplikativna konstanta linearne transformacije menjaju aritmetičku sredinu i standardnu devijaciju (kao što je objašnjeno u poglavlju **IV**, str 26 i 30):

$$(1)\ M_t = a + b*M$$
$$(2)\ S_t = b*S$$

U ovim jednačinama M i S predstavljaju aritmetičku sredinu i standardnu devijaciju izvornih rezultata, a M_t i S_t označavaju aritmetičku sredinu i standardnu devijaciju rezultata koji su iz izvornih dobijeni linearnom transformacijom. Budući da su nam u gornjim dvema jednačinama poznate aritmetičke sredina i standardna devijacija izvornih rezultata, a znamo i kakve treba da budu aritmetička sredina i standardna devijacija transformisanih rezultata, imamo dve nepoznate: a (aditivna konstanta) i b (multiplikativna konstanta). Iz jednačine (2) možemo odrediti rešenje za multiplikativnu konstantu:

$$b = \frac{S_t}{S}$$

a, zatim, ako u jednačini (1) konstantu b zamenimo izrazom $\frac{S_t}{S}$, možemo dobiti rešenje za aditivnu konstantu:

$$a = M_t - \frac{S_t}{S} * M$$

(prema Stockburger, 2000).

Ako želimo da, primenjujući dobijeno rešenje, rezultate na nekoj varijabli linearno transformišemo tako da aritmetička sredina transformisanih rezultata bude jednaka nuli, a da standardna devijacija ostane ista, primenili bismo linearnu transformaciju u kojoj bi multiplikativna konstanta b bila jednaka 1, dok bi aditivna konstanta a bila jednaka aritmetičkoj sredini izvornih rezultata sa negativnim predznakom, tj. a = 0 -1*M = -M. Zamenom ovih konstanti u linearnoj funkciji za transformaciju izvorne varijable x u transformisanu varijablu d dobijamo: d = x -M. Na taj način postižemo tzv. **centriranje** varijable.

Ukoliko aritmetička sredina transformisanih rezultata treba da bude jednaka nuli, a standardna devijacija jednaka jedinici, tada su neophodne vrednosti konstanti za transformaciju sledeće:

$$a = 0 - \frac{1}{S} * M = -\frac{M}{S}$$

$$b = \frac{1}{S}$$

Zamenom ovih konstanti u linearnoj funkciji za transformaciju izvorne varijable x u transformisanu varijablu z dobijamo:

$$z = -\frac{M}{S} + \frac{1}{S}x = \frac{x - M}{S}$$

Na taj način napravili smo transformaciju koja se uobičajeno zove **standardizacija** varijable x.

Na kraju, veoma čest slučaj u psihologiji je takav postupak linearnog transformisanja kojim se postiže da aritmetička sredina transformisanih rezultata bude 50, a standardna devijacija 10. U tom slučaju vrednosti konstanti za transformisanje su sledeće:

$$a = 50 - \frac{10}{S} * M$$

$$b = \frac{10}{S}$$

Na ovaj način dobijamo obrazac za pretvaranje izvornih rezultata na varijabli x u takozvane T-vrednosti:

$$T = (50 - \frac{10}{S} * M) + \frac{10}{S} x$$

Zapamtite:

- Linearne transformacije **ne menjaju oblik distribucije** rezultata. One, takodje, **ne menjaju nivo merne skale** sa koje potiču podaci. Na primer, ako izvorni podaci potiču sa intervalne merne skale i podaci koji su linearno transformisani zadržavaju svojstva intervalne skale.
- **I aditivna i multiplikativna** konstanta linearne transformacije **utiču na promenu aritmetičke sredine**, dok na **promenu standardne devijacije** ili varijanse utiče **samo multiplikativna, skalirajuća konstanta**.

****

Algebarske transformacije se u programu SPSS mogu izvesti korišćenjem komande **Compute** (koja je objašnjena u odeljku **IV.3.** na strani 23). U polje **Target Variable** ove komande upisuje se ime (**Name**) nove varijable u kojoj će biti sačuvane transformisane vrednosti, dok se transformacija definiše u polju **Numeric Expression**.

Primeri:

1. Izvorna varijabla EPQN sadrži broj poena, tj. skor na varijabli ekstraverzija. Izračunali smo aritmetičku sredinu i standardnu devijaciju izvornih rezultata i one iznose 15 i 5. U novoj varijabli C_EPQN treba da budu centrirani rezultati ili devijacioni skorovi ispitanika.
U polje **Target Variable** upisaćemo **C_EPQN**, a u polje **Numeric Expression** upisaćemo **EPQN-15**
Kada kliknemo na dugme **OK**, program će napraviti novu varijablu na kraju (tj. "desnom kraju") aktivnog fajla sa podacima.

2. Želimo da standardizujemo rezultate na varijabli EPQN. U polje **Target Variable** upisaćemo **Z_EPQN**, a u polje **Numeric Expression** upisaćemo **(EPQN-15)/5**

3. Želimo da pretvorimo izvorne podatke iz varijable BRGR koji sadrže broj grešaka u eksperimentu korišćenjem kvadratno-korenske transformacije. U polje **Target Variable** upisaćemo **KK_BRGR**, a u polje **Numeric Expression** ubacićemo iz spiska **Functions** funkciju **SQRT(?)** i zameniti znak pitanja imenom izvorne varijable, u ovom slučaju **BRGR**.

4. Želimo da pretvorimo izvorne podatke iz varijable VR koja sadrži vremena reagovanja ispitanika u eksperimentu korišćenjem logaritamske transformacije. U polje **Target Variable** upisaćemo **L_VR**, a u polje **Numeric Expression** ubacićemo iz spiska **Functions** funkciju **LN(?)** i zameniti znak pitanja imenom izvorne varijable, u ovom slučaju **VR**.

5. Želimo da podatke na varijabli EPQE, koja sadrži izvorne rezultate sa upitnika za varijablu ekstraverzija, pretvorimo u skalu čija će aritmetička sredina biti 50, a standardna devijacija 10. Izračunali smo prethodno da je aritmetička sredina izvorne varijable EPQE 14, a standardna devijacija 2. U polje **Target Variable** upisaćemo **T_EPQE**, a u polje **Numeric Expression** upisaćemo **(10/2)*EPQE + (50 - 10*14/2)** ili, jednostavnije, **5*EPQE -20**.

****

2. Površinske transformacije sastoje se u prevodjenju sirovih rezultata u transformisane vrednosti koje govore o tome koliki deo od ukupne "površine" distribucije (kada se distribucija grafički prikaže) je ispod ili iznat date vrednosti, tj. koliki procenat rezultata u distribuciji leži ispod ili iznad tako dobijene vrednosti. Najpoznatije površinske transformacije u psihologiji su prevodjenje sirovih rezultata u percentilne (ili, pak, decilne ili kvartilne) i normalizovane vrednosti. Percentilna, decilna ili kvartilna vrednost koja odgovara sirovom rezultatu ispitanika pokazuje kojem percentilu, decilu, ili kvartilu u distribuciji taj rezultat pripada (videti odeljak *IV.6. Percentili i percentilni rangovi*). Normalizovane vrednosti koje odgovaraju određenom izvornom rezultatu dobijaju se tako što se izračunaju kumulativne relativne frekvence (kumulativne frekvence u procentima ili kumulativne proporcije) za svaki sirovi rezultat i zatim se sirovom rezultatu pridružuje onaj normalizovani rezultat koji na normalnoj raspodeli odgovara njegovoj kumulativnoj relativnoj frekvenci. (O Normalnoj raspodeli ili funkciji gustine može se pročitati na strani 19 Dodatka 2) . Pri prevodjenju izvornih rezultata u normalizovane vrednosti treba imati na umu da ovakva transformacija ima smisla **samo ako se može pretpostaviti** da se varijabla čiji su to rezultati normalno ditribuira u populaciji i ako možemo biti uvereni da su aritmetička sredina i standardna devijacija koje smo dobili na uzorku dobra ocena populacijske aritmetičke sredine i standardne devijacije (ocenjivanje parametara objašnjeno je u odeljku *VI.1. Statističko ocenjivanje parametara*, str. 49 do 55)

****

Površinske transformacije u programu SPSS moguće je izvesti korišćenjem procedure **Transform/Rank Cases**. U polje **Variable(s)** prozora za dijalog **Rank Cases** ubaci se ime varijable koja sadrži izvorne podatke, zatim klikne na dugme **Rank Types.** U novom prozoru **Rank Cases: Types** treba uključiti opciju **Ntiles** (klikom na beli kvadratić ispred teksta Ntiles) i u ulazno polje posle reči Ntiles upisati **4** (za kvartile), **10** (za decile) ili **99** (za percentile). Program će automatski napraviti novu varijablu koja će u svom imenu imati slova **nti** ispred nekog broja (zavisno od toga koliko takvih varijabli već ima u aktivnom fajlu) i u takvu varijablu upisaće percentilne (decilne, kvartilne) vrednosti.

Pretvaranje izvornih rezultata u normalizovane skorove može se izvesti korišćenjem procedure **Transform/Rank Cases**. U polje **Variable(s)** prozora za dijalog **Rank Cases** ubaci se ime varijable koja sadrži izvorne podatke, zatim klikne

na dugme **Rank Types.** U novom prozoru **Rank Cases: Types** treba kliknuti na dugme **More** i uključiti opciju **Normal Scores**. Uobičajenu opciju (koja je automatski uključena u okviru **Proportion Estimation Formula**) treba menjati samo kada smo sigurni da Blumov postupak želimo da zamenimo nekim od preostalih načina ocene kumulativne proporcije u distribuciji koja odgovara nekom rangu. Program će automatski napraviti novu varijablu koja će u svom imenu imati ispred imena izvorne varijable slovo **n** ili, pak, slova **nor** ispred nekog broja (zavisno od toga koliko takvih varijabli već ima u aktivnom fajlu) i u takvu varijablu upisaće normalizovane vrednosti.

****

V. STATISTIČKI OPIS I GRAFIČKI PRIKAZ JEDNE KATEGORIČKE (NOMINALNE) VARIJABLE

10. (str. 45) Entropija nominalne varijable

Razumevanje obrazaca za entropiju podrazumeva poznavanje nekih osnovnih osobina logaritamske funkcije:
Logaritamska funkcija ima oblik:

$$f(x) = \log_a x,\ x \in \mathbf{R}_+,\ a > 0 \text{ i } a \neq 1, \text{ pri čemu je } \mathbf{R}_+ = (0, \infty).$$

Ako je a = e = 2. 71828, pri čemu je e Ojlerova konstanta [transcendentni broj definisan kao $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$], tada se radi o tzv. **prirodnom** logaritmu, u oznaci ln x.
O logaritmima je korisno misliti na sledeći način: logaritam od x za osnovu a je eksponent na koji treba podići osnovu a da bi se dobila vrednost x, tj. $a^{f(x)} = x$. Tada je $f(x)$ logaritam za x, dok je x antilogaritam. Npr. osnovu 2 treba dići na stepen 3 da bi se dobilo 8, pa je $\log_2 8 = 3$. U tom slučaju 8 je antilogaritam.

Ovde ćemo pod logaritmom podrazumevati prirodni logaritam.

Osnovne osobine logaritama:

1. ln 0 ne postoji. Nijedan broj dignut na osnovu e ne može dati nulu!
2. ln 1 = 0.
3. Ako je 0 < x < 1, ln x < 0, a ako je x > 1, ln x > 0.
4. Funkcija $f(x)$ = ln x monotono raste jer je osnova e > 1.
5. ln x*y = ln x + ln y.
6. ln x/y = ln x - ln y.
7. $\ln x^z = z*\ln x$.
8. $e^{\ln x} = x$.
9. $\ln e^x = x$

XIV. MULTIPLA REGRESIONA ANALIZA

11. (Dodatak za kraj poglavlja, str. 168)

XIV.4. Medijatorske, moderatorske i supresorske prediktorske varijable u regresionoj analizi

Pri izboru prediktorskih varijabli, kao i pri tumačenju rezultata multiple regresione analize, veoma je važno, iz teorijskih i praktičnih razloga, prepoznati medju potencijalnim ili već uključenim prediktorima one prediktore koji su **medijatori**, **moderatori** ili, pak, **supresori**.

Kriterijumi za **medijatorsku** ulogu prediktorske varijable u regresionom modelu

1. Prediktorska varijabla koja nije medijator je u značajnoj korelaciji sa kriterijumskom;
2. Prediktorska varijabla koja nije medijator je u značajnoj korelaciji sa medijatorskom prediktorskom varijablom;
3. Medijatorska prediktorska varijabla je u značajnoj korelaciji sa kriterijumskom;
4. Kada se statistički kontroliše medijatorska varijabla, postojeća korelacija prediktorske varijable sa kriterijumskom varijablom se smanjuje ili nije više statistički značajna. Drugim rečima, kada se statistički ispitanici izjednače po medijatorskoj prediktorskoj varijabli, parcijalna korelacija između prediktorske nemedijatorske i kriterijumske varijable jeste jako niska, tj. blizu nule.
   Prepoznavanje medijatorske uloge neke prediktorske varijable veoma je korisno u teorijskom smislu jer pomaže potpunijem razjašnjenju mehanizama determinisanja individualnih razlika medju ljudima u pogledu određene kriterijumske aktivnosti.

Moderatorska varijabla mora biti u interakciji sa nekom od preostalih prediktorskih varijabli u predviđanju kriterijumske varijable, premda sama ima zanemarljivu korelaciju sa kriterijumskom varijablom. Dakle, prediktor koji igra moderatorsku ulogu u regresionom modelu, mada sam ne doprinosi objašnjenju ili predviđanju individualnih razlika u pogledu kriterijumske varijable, u interakciji sa nekom od preostalih prediktorskih varijabli bitno doprinosi objašnjenju kriterijumske varijable.

Supresorska varijabla (klasični tip supresora)

Ukoliko je u regresioni model uključen prediktor koji sam ne korelira sa kriterijumskom varijablom, ali se na ispisu iz regresione analize može uočiti da je regresioni koeficijent za ovaj prediktor negativan i znatno odstupa od nule, potrebno je proveriti da li se možda radi o tzv. supresorskoj varijabli. Kriterijumi koji pokazuju da li neki prediktor, koji je uključen u regresioni model zajedno sa ostalim prediktorima, spada u tzv. klasični tip supresora su sledeći:

1. Supresorska prediktorska varijabla ne korelira sa kriterijumskom varijablom, ali korelira sa nekim drugim prediktorom za koji je ona supresor;
2. Prediktorska varijabla koja nije supresor, ali na koju supresor deluje, korelira sa kriterijumskom varijablom;

3. Parcijalna korelacija između supresorske i kriterijumske varijable je negativna;
4. Parcijalni standardizovani regresioni koeficijenti su veći (čak mogu biti i veći od maksimalno moguće korelacije, tj. od 1) i za prediktorsku i za supresorsku varijablu od standardizovanog koeficijenta iz jednostruke regresione analize, a regresioni koeficijent supresorske varijable je negativan.

Dakle, uprkos tome što ne korelira sa kriterijumskom varijablom, supresorska varijabla vrši "supresiju" određenog dela varijanse druge prediktorske varijable u regresionom modelu i to onog dela varijanse koji je irelevantan za predviđanje kriterijumske varijable. Na taj način, kvadrat multiple determinacije, tj. procenat objašnjene varijanse kriterijumske varijable postaje viši no u modelu bez supresora. Podrobnije objašnjenje i analiza različitih vrsta supresorskog efekta prediktora u regresionom modelu dati su u članku Conger-a (Conger, 1974).

16. Milligan, G.W. & Cooper, M.C. (1985). An Examination of Procedures for determining the Number of Clusters in a Data Set, *Psychometrika*, 50,159-179.

17. Momirović, K. (1988). *Uvod u analizu nominalnih varijabli*, Metodološke sveske 2, Ljubljana: JUS.

18. Momirović, K., Wolf, B. & Popović, D.A. (1999). *Uvod u teoriju merenja I: Interne metrijske karakteristike kompozitnih mernih instrumenata*, Priština: Univerzitet u Prištini, Fakultet za fizičku kulturu.

19. Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*, Second Edition, Belmont: Duxbury Press.

20. Stilson D. W. (1966). *Probability and Statistics in psychological research and theory*, San Francisko: Holden-Day, Inc.

21. Stockburger, D.W. (2000). Linear transformations, www dokument, URL: http://www.psychstat.smsu.edu/introbook/sbk15m.htm.

22. Vuković, N. (1997). *PC Verovatnoća i statistika*, Beograd:Fakultet organizacionih nauka.

23. Zorić, A. (1996). *Faktorska analiza binarnih varijabli u generalizovanom imaž prostoru*, Diplomski rad, Beograd:Filozofski fakultet.

24. Wilks, S.S. (1962). *Mathematical Statistics*, New York: John Wiley & Sons, Inc.

25. Winer, B.J. (1970). *Statistical Principles in Experimental Design*, New-York: McGraw-Hill.

26. Winkler, R.W. & Hays, W.L. (1975). *Statistics Probability, inference and decision* (second edition), New York: Holt, Rinehart and Winston.

Literatura:

1. Conger, A.J. (1974). A revised definition for suppressor variables: A Guide to their identification and interpretation, *Educational and Psychological Measurement, 34*, 35-46.

2. Deming, V.E. (1952). *Teorija reprezentativne metode*, Beograd: Izdavačko preduzeće Saveznog zavoda za statistiku i evidenciju.

3. Dragićević, Č. (1984). *Statistika za psihologe*, Beograd: Savez društava psihologa Srbije.

4. Everit, B.S. (1996). *Making sense of Statistics in Psychology*, Oxford: Oxford University Press.

5. Ferguson, G.A. (1981). *Statistical Analysis in Psychology and Education*, Fifth edition, Auckland: McGRAW-HILL INTERNATIONAL BOOK COMPANY

6. Ferligoj, A. (1989). *Razvrščanje v skupine* (na slovenačkom, Razvrstavanje u grupe), Ljubljana: JUS.

7. Гихман И.И., Скороход, А.В. & Ядренко, М.И. (1988). *Теория вероятностей и математическая статистика*, Киев: "Вища школа".

8. Green, P.E & Carroll, J.D. (1976). *Mathematical tools for applied multivariate analysis*, London: Academic Press.

9. Horst, P. (1966). Psychological Measurement and Prediction, Belmont: Wadsworth Publishing Company, Inc.

10. Howell, D.C. (1997). *Statistical Methods for Psychology*, 4th ed.Belmont:Duxbury Press.

11. Ivković, Z. (1990). *Matematička statistika*, Beograd: Naučna knjiga.

12. Ivanović, B. (1966). *Teorijska statistika*, Beograd: Jugoslovenski institut za ekonomska istraživanja;

13. Kornbrot, D.E. (1990).The Rank Difference Test: A new and meaningful alternative to the Wilcoxon signed ranks test for ordinal data, *British Journal of Mathematical and Statistical Psychology*, 43, 241-264

14. Kovačić, Z. (1994). *Multivarijaciona analiza*, Beograd: Ekonomski fakultet.

15. Michell, J. (1990). *An introduction to the logic of psychological measurement*, Hillsdale, NJ:Erbaum, 1990.

Lazar Tenjović

STATISTIKA U PSIHOLOGIJI
Priručnik
Dopunjeno izdanje

Izdavač
Centar za primenjenu psihologiju
Beograd, Đušina br. 7/III
tel./faks: 3239-685

Za izdavača
Duško Babic

Likovno rešenje korica
Dragana Božić

Tiraž: 500 primeraka

Štampa
Centar za primenjenu psihologiju, Beograd

ISBN 86-83797-29-5

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

159 . 9 : 519 . 2 (035)

ТЕЊОВИЋ, Лазар
Statistika u psihologiji : priručnik /
Lazar Tenjović. – Beograd : Centar za primenjenu psihologiju Društva psihologa Srbije, 2000 (Beograd : Centar za primenjenu psihologiju Društva psihologa Srbije). – 1 knj. (razl. pag.) : tabele, graf. prikazi : 24 cm

Tiraž 500. – Bibliografija

159 . 9 . 072 . 075 (035) 519 . 688 : 159 . 9 (035)

а) Психологија – Статистички методи – Приручници б) Апликативни програм "SPSS for Windows"